AD_____

Award Number: DAMD17-98-2-8005

TITLE: Malaria Genome Sequencing Project

PRINCIPAL INVESTIGATOR: Malcolm J. Gardner, Ph.D.

CONTRACTING ORGANIZATION: The Institute for Genomic Research
Rockville, Maryland 20850

REPORT DATE: January 2003

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 074-0188*



| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | January 2003 | Annual (17 Dec 01 - 16 Dec 02) |

**4. TITLE AND SUBTITLE**
Malaria Genome Sequencing Project

**5. FUNDING NUMBERS**
DAMD17-98-2-8005

**6. AUTHOR(S) :**
Malcolm J. Gardner, Ph.D.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
The Institute for Genomic Research
Rockville, Maryland 20850

Email: gardner@tigr.org

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES**
Original contains color plates: All DTIC reproductions will be in black and white.

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**12b. DISTRIBUTION CODE**

**13. Abstract** ***(Maximum 200 Words)*** *(abstract should contain no proprietary or confidential information)*
The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims were added to the Cooperative Agreement: **Specific Aim 4**, sequencing of *P. yoelii* to 5X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 5X coverage. This year we reached a major milestone by publishing, in collaboration with the Sanger Institute and Stanford University, the complete genome sequence of *P. falciparum* in the journal *Nature*. In addition, we published a comparative analysis of the genome of the rodent malaria parasite *P. yoelii* with that of *P. falciparum*. We began sequencing of the second major human malaria parasite P. vivax and attained 5X coverage of the genome. We obtained additional funds from other sources to permit the sequencing of *P. vivax* to 8X coverage, to close one-third of the genome, and to annotate the genome and compare it to the genomes of *P. falciparum* and *P. yoelii*. This work will be completed under a 12-month no-cost extension of this Cooperative Agreement. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

**14. SUBJECT TERMS:**
P. falciparum, P. vivax, P. yoelii, malaria, genome, chromosome

**15. NUMBER OF PAGES**
51

**16. PRICE CODE**

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | Unlimited |

**Table of Contents**

## Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many tropical areas of the world, and also affects many individuals and military forces that visit these areas. In 1994 the World Health Organization estimated that there were 300-500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably. These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control [1]. Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism [2], and establishes an excellent starting point for this process. In 1995, an international consortium including the National Institutes of Health, the Wellcome Trust, the Burroughs Wellcome Fund, and the US Department of Defense was formed (Malaria Genome Sequencing Project) to finance and coordinate genome sequencing of the human malaria parasite *Plasmodium falciparum*, and later, a second, yet to be determined, species of *Plasmodium*. Another major goal of the consortium was to foster close collaboration between members of the consortium and other agencies such as the World Health Organization, so that the knowledge generated by the Project could be rapidly applied to basic research and antimalarial drug and vaccine development programs worldwide. Participating centers include the Naval Medical Research Center, the Wellcome Trust Sanger Institute, and the Stanford University Genome Technology Center.

## Body

This report describes progress in the Malaria Genome Sequencing Project achieved by The Institute for Genomic Research and the Malaria Program, Naval Medical Research Center, under Cooperative Research Agreement DAMD17-98-2-8005, over the 12 month period from Dec. '01 to Dec '02. The Specific Aims of the work supported by this agreement are listed below. Specific Aims 1-3 were contained in the original Cooperative Agreement. Specific Aims 4-5 were added to the Cooperative Agreement through modifications.

The Cooperative Agreement was initially scheduled to expire in December 2002. However, we were recently granted a 12-month no-cost extension to allow us to complete a newly-expanded Specific Aim 5 (Sequencing of *P. vivax* to 5X coverage).

**1. Determine the sequence of 3.5 megabases of the *P. falciparum* genome (clone 3D7):**

a) Construct small-insert shotgun libraries (1-2 kb inserts) of chromosomal DNA isolated from preparative pulsed-field gels.

b) Sequence a sufficiently large number of randomly selected clones from a shotgun library to provide 10-fold coverage of the selected chromosome.

c) Construct P1 artificial chromosome (PAC) libraries (inserts up to 20 kb) of chromosomal DNA isolated from preparative pulsed-field gels.

d) If necessary, generate additional STS markers for the chromosome by i) mapping unique-sequence contigs derived from assembly of the random sequences to chromosome, ii) mapping end-sequences from chromosome-specific PAC clones to YACs.

e) Use TIGR Assembler to assemble random sequence fragments, and order contigs by comparison to the STS markers on each chromosome.

f) Close any remaining gaps in the chromosome sequence by PCR and primer-walking using *P. falciparum* genomic DNA or the YAC, BAC, or PAC clones from each chromosome as templates.

**2. Analyze and annotate the genome sequence:**

a) employ a variety of computer techniques to predict gene structures and relate them to known proteins by similarity searches against databases; identify untranslated features such as tRNA genes, rRNA genes, insertion sequences and repetitive elements; determine potential regulatory sequences and ribosome binding sites; use these data to identify metabolic pathways in *P. falciparum.*

**3. Establish a publicly-accessible *P. falciparum* genome database and submit sequences to GenBank.**

**4. Perform whole genome shotgun sequencing of the rodent malaria parasite *Plasmodium yoelii* to 3X coverage, assemble into contigs, annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.**

**5. Perform whole genome shotgun sequencing of the human malaria parasite *Plasmodium vivax* to 5X coverage, assemble the contigs, annotate the**

**contigs, make the data available on the TIGR web site, and submit the data to GenBank.**

We are pleased to report that excellent progress has been made towards achievement of these goals. In previous annual reports we announced the publication in *Science* of the first complete sequence of a malarial chromosome (chromosome 2) [3]; development of a *Plasmodium* gene finding program, GlimmerM [4]; introduction of optical restriction mapping technology for rapid mapping of whole *Plasmodium* chromosomes [5,6]; completion of the random phase of sequencing of 3 additional *P. falciparum* chromosomes and major progress in gap closure; sequencing of the rodent malaria parasite *Plasmodium yoelii* to 5X coverage and release of preliminary annotation of this genome on the TIGR web site (http://www.tigr.org/tdb/edb2/pya1/htmls/). Through a subcontract to Dr. John Yates at the Scripps Institute, we also assisted NMRC in their pilot project to apply the techniques of proteomics towards the identification of novel antigens in parasite (sporozoite) extracts. Finally, we have continually reviewed with NMRC further steps that can be taken to more rapidly apply *Plasmodium* genomics, functional genomics, and proteomics to problems of vaccine development for malaria.

Over the past year, we have completed specific Aims 1, 2, 3, and 4. Chromosomes 10, 11, and 14 have been completed, and in collaboration with the Sanger Institute, Stanford University, and NMRC, the complete *P. falciparum* genome sequence was published in *Nature* on Oct. 3, 2002 [7,8].We also completed the sequencing of *P. yoelii* to 5X coverage, and published this sequence and a comparison to *P. falciparum* [9]. Finally, we have continued to sequence the genome of *P. vivax* and have attained 6X coverage. This data was released to the public on the TIGR website.

**Sequencing of *P. falciparum* chromosomes 10, 11, and 14 (Specific Aims 1, 2, 3)**

Sequencing of chromosome 10, 11, and 14 was funded primarily by grants from the NIAID (chromosomes 10 and 11) and the Burroughs Wellcome Fund (chromosome 14). Funds from this collaborative agreement were used to accelerate the sequencing, assist in closure and annotation, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. In previous years we described the isolation of chromosomal DNA, preparation of shotgun libraries, random sequencing, assembly, gap closure, production and public release of preliminary annotation. This past year focused primarily on gap closure and the final annotation and publication of the *P. falciparum* genome sequence in collaboration with the other members of the *P. falciparum* genome consortium.

All gaps in chromosomes 10, 11, and 14 have now been closed. In the next few months we will be submitting the revised sequences and updated annotation to GenBank and PlasmoDB.

**Annotation and publication of the *P. falciparum* genome sequence (Specific Aims 2 and 3)**

In last year's report we described an agreement made with the Sanger Institute and Stanford University to collaborate on the joint analysis and publication of entire *P. falciparum* genome sequence. This whole genome overview was to be accompanied by a series of papers by each sequencing center on the chromosomes sequenced by each group. The whole genome overview and chromosome papers were to be published in a single issue of a journal. In addition, a comparative analysis of the *P. falciparum* and *P. yoelii* genomes based upon the 5X coverage *P. yoelii* sequence was to be be published along with the P. *falciparum* papers.

The principal investigator of this agreement was selected to be the coordinator of the annotation effort and the lead author on the final publication. Furthermore, TIGR was chosen to be the central repository of all the *P. falciparum* genome data. Over the past year, TIGR collected the chromosome sequences and associated annotation from the other sequencing centers and coordinated the analysis of the genome sequence and the preparation of whole genome and a series of chromosome manuscripts for publication. The manuscripts were submitted for publication in July 2002 and published in *Nature* on Oct. 3, 2002 [7,8].

**Sequencing of *P. yoelii* to 5X coverage (Specific Aim 4)**

A secondary goal established at the initiation of the malaria genome project was to sequence the genome of a another species of *Plasmodium* so as to be able to perform a series of comparative analyses.

After discussions with NMRC we elected to proceed with sequencing of *P. yoelii*. Reductions in the costs of sequencing allowed us to perform this work without requesting additional funds. The genome was sequenced to 5X coverage and a comparative analysis with the *P. falciparum* genome was performed. This work was published in *Nature* on Oct. 3, 2002 [9].

**Sequencing of *P. vivax* to 5X coverage (Specific Aim 5)**

*P. vivax* is the second most important human malaria parasite. In last year's report, we described the addition of this Specific Aim to the Cooperative Agreement. Two genomic shotgun libraries of P. vivax (Sall strain) were constructed using DNA provided by John Barnwell of the Centers for Disease Control. One library contained 2-3 kb inserts and the other library contained 5-6 kb inserts. As of Dec. 15, 2002, 220,7763 sequences with an average read length of 650 nucleotides have been generated at a success rate of 85.6%. A preliminary assembly has been performed which suggests

that the *P. vivax* genome is about 23 Mb in size, about the same as that of *P. falciparum*. Furthermore, the use of libraries with inserts of 5-6 kb resulted in the generation of large scaffolds of contigs linked by forward-reverse read pairs, which will faciliate gap closure. Preliminary contigs are available for searching on the TIGR web site (http://www.tigr.org/tdb/e2k1/pva1/).

Our original intention was to sequence to 5X coverage and perform a comparative analysis with *P. falciparum*. However, due to reductions in sequencing costs and additional funds obtained from a no-cost extension to an NIAID-funded project, we should be able to obtain 8X coverage and close up to one-third of the *P. vivax* genome. This will facilitate a more detailed analysis of the genome sequence and assist in the identification of novel vaccine and drug targets for this relatively little-studied malaria parasite.

To summarize the sequencing efforts in Specific Aims 1-5, by the end of this cooperative agreement, the complete genome sequence of *P. falciparum* will have been published, as will a comparative analysis of the *P. falciparum* and *P. yoelii* genomes. The *P. vivax* genome will have been sequenced to 5X coverage. A 3-way comparative analysis of the *P. falciparum, P. vivax*, and *P. yoelii* genomes will be also be performed, but this is unlikely to be completed until after this cooperative agreement has expired.

**Proteomics studies**

A major goal of the malaria genome project is to identify antigens for vaccine development. Analysis of the genome sequence data can be used to identify potential antigens but does not by itself provide all of the information required for selection and prioritization of vaccine candidates. For example, the genome sequence itself does not specify at which point in the life cycle a gene is transcribed, or whether the protein product of a gene is actually present in the parasite. To identify proteins present in various stages of the parasite life cycle, we have begun to use proteomics techniques to directly identify parasite proteins in cell lysates.

In the last two annual reports we reported on work done by Dr. John Yates and colleagues at the Scripps Research Institute, partly funded by a subcontract from TIGR under this cooperative agreement. Briefly, proteins in parasite lysates were digested with proteases and the resulting peptides were separated by high-resolution liquid chromatography. The peptides were then injected into a tandem mass spectrometer. Spectra of each peptide were matched against predicted spectra of the peptides predicted from the genome sequence. In this way peptides generated from cell lysates were used to identify the proteins present in the cell lysate. Our role has been mainly to provide Dr. Yates's group with genomic sequence data from *P. falciparum* and *P. yoelii*, which they used to identify peptides derived from parasite lysates. Over 2,400 *P. falciparum* proteins, about 45% of the total proteins predicted from the genome sequence, were identified, including approx 500 proteins from sporozoite stages[10]. The NMRC is using this data to select antigens for vaccine development.

## Key Research Accomplishments

1) The sequences of chromosomes 2, 10, 11, and 14 were completed.

2) Chromosomes 2, 10, 11, and 14 were annotated at TIGR.

3) In collaboration with the Sanger Institute and Stanford University, the entire *P. falciparum* genome was annotated.

4) The *P. yoelii* genome sequence obtained at 5X coverage was annotated.

5) Sequencing of the *P. vivax* genome reached 6X coverage. Preliminary assemblies were performed and indicate that the final sequence will be of high quality.

## Reportable Outcomes

1) Web site. Preliminary contigs and annotation for the *P. vivax* genome at 5X coverage. (http://www.tigr.org/tdb/e2k1/pva1)).

2) Web site. Final annotation of the *P. falciparum* genome (http://www.tigr.org/tdb/e2k1/pfa1/

3) Publication. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

4) Publication. Gardner, M. J., Shallom, S., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C. Y., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter,

J. C., Carucci, D. J., Hoffman, S. L. & Fraser, C. M. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. *Nature* **419**, 531-534 (2002).

5) Publication. Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T., Pertea, M., Ermolaeva, M. D., Allen, D. R., Silva, J. C., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Shoaibi, A., Cummings, L. M., Cho, J. K., Quackenbush, J., van Aken, S. E., Riedmuller, S. B., Feldbylum, T. V., Florens, L., Yates III, J. R., Raine, D. J., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419**, 512-519 (2002).

6) M. J. Gardner, Complete genome sequence of the human malaria parasite *Plasmodium falciparum*. Press conference to announce publication of the *P. falciparum* genome sequence in *Nature*. Held at the American Association for the Advancement of Science, Washington, D.C., Oct. 3rd, 2002.

7) M. J. Gardner, Complete genome sequence of *Plasmodium falciparum*. Paper presented at the American Society of Tropical Medicine and Hygiene Annual Meeting, Denver, Nov. 2002.

8) M. J. Gardner, Complete genome sequence of *Plasmodium falciparum*. Paper presented at the 3rd Multilateral Initiative on Malaria Meeting, Arusha, Tanzania, Nov. 2002.

**Conclusions**

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims were added to the Cooperative Agreement: **Specific Aim 4**, sequencing of *P. yoelii* to 3X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 3X coverage.

By publishing the complete genome sequence of *P. falciparum*, and chromosomes 2, 10, 11 and 14, we have **completed Specific Aims 1-3.** By sequencing *P. yoelii* to 5X coverage and publishing an analysis of this genome we have **completed Specific Aim 4.** We received a 12-month no-cost extension to assist in completion of **Specific Aim 5**, sequencing of *P. vivax* to 5X coverage. This has been completed, but by using funds from other sources we will be able to enhance Specific Aim 5 by closing up to one-third of the *P. vivax* genome sequence, annotating the

genome sequence, and perform a comparative analysis of the *P. vivax* and *P. falciparum* genomes.

## References


1. Butler, D., Maurice, J. & O'Brien, C. Briefing malaria. *Nature* **386**, 535-540 (1997).
2. Bloom, B. R. A microbial minimalist. *Nature* **378**, 236 (1995).
3. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pedersen, J., Shen, K., Jing, J., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G. G., Clayton, R. L., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C. & Hoffman, S. L. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132 (1998).
4. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24-31 (1999).
5. Jing, J., Aston, C., Zhongwu, L., Carucci, D. J., Gardner, M. J., Venter, J. C. & Schwartz, D. C. Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Research* **9**, 175-181 (1999).
6. Lai, Z., Jing, J., Aston, C., Clarke, V., Apodaca, J., Dimlanta, E. T., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T., Paxia, S., Hoffman, S. L., Venter, J. C., Huff, E. J. & Schwartz, D. C. A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genetics* **23**, 309-313 (1999).
7. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).
8. Gardner, M. J., Shallom, S., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C. Y., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L. & Fraser, C. M. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. *Nature* **419**, 531-534 (2002).
9. Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T., Pertea, M., Ermolaeva, M. D., Allen, D. R., Silva, J. C., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Shoaibi, A., Cummings, L. M., Cho, J. K., Quackenbush, J., van Aken, S. E., Riedmuller, S. B., Feldbylum, T. V., Florens, L., Yates III, J. R., Raine, D. J., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B.,

van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419**, 512-519 (2002).

10. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., 3rd & Carucci, D. J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520-526 (2002).

## Appendices

Appendix A. Reprint: Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. & Barrell, B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

Appendix B. Reprint: Gardner, M. J., Shallom, S., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C. Y., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L. & Fraser, C. M. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11, and 14. *Nature* **419**, 531-534 (2002).

Appendix C. Reprint: Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T., Pertea, M., Ermolaeva, M. D., Allen, D. R., Silva, J. C., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., Shoaibi, A., Cummings, L. M., Cho, J. K., Quackenbush, J., van Aken, S. E., Riedmuller, S. B., Feldbylum, T. V., Florens, L., Yates III, J. R., Raine, D. J., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. & Carucci, D. J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419**, 512-519 (2002).

Appendix D. Reprint: Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., 3rd & Carucci, D. J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520-526 (2002).

articles

# Genome sequence of the human malaria parasite *Plasmodium falciparum*

Malcolm J. Gardner[1], Neil Hall[2], Eula Fung[3], Owen White[1], Matthew Berriman[2], Richard W. Hyman[3], Jane M. Carlton[1], Arnab Pain[2], Karen E. Nelson[1], Sharen Bowman[2]*, Ian T. Paulsen[1], Keith James[2], Jonathan A. Eisen[1], Kim Rutherford[2], Steven L. Salzberg[1], Alister Craig[4], Sue Kyes[5], Man-Suen Chan[5], Vishvanath Nene[1], Shamira J. Shallom[1], Bernard Suh[1], Jeremy Peterson[1], Sam Angiuoli[1], Mihaela Pertea[1], Jonathan Allen[1], Jeremy Selengut[1], Daniel Haft[1], Michael W. Mather[6], Akhil B. Vaidya[6], David M. A. Martin[7], Alan H. Fairlamb[7], Martin J. Fraunholz[8], David S. Roos[8], Stuart A. Ralph[9], Geoffrey I. McFadden[9], Leda M. Cummings[1], G. Mani Subramanian[10], Chris Mungall[11], J. Craig Venter[12], Daniel J. Carucci[13], Stephen L. Hoffman[13]*, Chris Newbold[5], Ronald W. Davis[3], Claire M. Fraser[1] & Bart Barrell[2]

**The parasite *Plasmodium falciparum* is responsible for hundreds of millions of cases of malaria, and kills more than one million African children annually. Here we report an analysis of the genome sequence of *P. falciparum* clone 3D7. The 23-megabase nuclear genome consists of 14 chromosomes, encodes about 5,300 genes, and is the most (A + T)-rich genome sequenced to date. Genes involved in antigenic variation are concentrated in the subtelomeric regions of the chromosomes. Compared to the genomes of free-living eukaryotic microbes, the genome of this intracellular parasite encodes fewer enzymes and transporters, but a large proportion of genes are devoted to immune evasion and host–parasite interactions. Many nuclear-encoded proteins are targeted to the apicoplast, an organelle involved in fatty-acid and isoprenoid metabolism. The genome sequence provides the foundation for future studies of this organism, and is being exploited in the search for new drugs and vaccines to fight malaria.**

Despite more than a century of efforts to eradicate or control malaria, the disease remains a major and growing threat to the public health and economic development of countries in the tropical and subtropical regions of the world. Approximately 40% of the world's population lives in areas where malaria is transmitted. There are an estimated 300–500 million cases and up to 2.7 million deaths from malaria each year. The mortality levels are greatest in sub-Saharan Africa, where children under 5 years of age account for 90% of all deaths due to malaria[1]. Human malaria is caused by infection with intracellular parasites of the genus *Plasmodium* that are transmitted by *Anopheles* mosquitoes. Of the four species of *Plasmodium* that infect humans, *Plasmodium falciparum* is the most lethal. Resistance to anti-malarial drugs and insecticides, the decay of public health infrastructure, population movements, political unrest, and environmental changes are contributing to the spread of malaria[2]. In countries with endemic malaria, the annual economic growth rates over a 25-year period were 1.5% lower than in other countries. This implies that the cumulative effect of the lower annual economic output in a malaria-endemic country was a 50% reduction in the per capita GDP compared to a non-malarious country[3]. Recent studies suggest that the number of malaria cases may double in 20 years if new methods of control are not devised and implemented[1].

An international effort[4] was launched in 1996 to sequence the *P. falciparum* genome with the expectation that the genome sequence would open new avenues for research. The sequences of two of the 14 chromosomes, representing 8% of the nuclear genome, were published previously[5,6] and the accompanying Letters in this issue describe the sequences of chromosomes 1, 3–9 and 13 (ref. 7), 2, 10, 11 and 14 (ref. 8), and 12 (ref. 9). Here we report an analysis of the genome sequence of *P. falciparum* clone 3D7, including descriptions of chromosome structure, gene content, functional classification of proteins, metabolism and transport, and other features of parasite biology.

## Sequencing strategy

A whole chromosome shotgun sequencing strategy was used to determine the genome sequence of *P. falciparum* clone 3D7. This approach was taken because a whole genome shotgun strategy was not feasible or cost-effective with the technology that was available at the beginning of the project. Also, high-quality large insert libraries of (A + T)-rich *P. falciparum* DNA have never been constructed in *Escherichia coli*, which ruled out a clone-by-clone sequencing strategy. The chromosomes were separated on pulsed field gels, and chromosomal DNA was extracted and used to construct shotgun libraries of 1–3-kilobase (kb) fragments of sheared DNA. Eleven of the fourteen chromosomes could be resolved on the gels, but chromosomes 6, 7 and 8 could not be resolved and were sequenced as a group. The shotgun sequences were assembled into contiguous DNA sequences (contigs), in some cases with low coverage shotgun sequences of yeast artificial chromosome (YAC) clones to assist in the ordering of contigs for closure. Sequence tagged sites (STSs)[10], microsatellite markers[11,12] and HAPPY mapping[7] were also used to place and orient contigs during the gap closure process. The high (A + T) content of the genome made gap closure extremely difficult[7–9]. The predicted restriction enzyme maps of the chromosome sequences were compared to optical restriction maps to verify that the chromosomes had been assembled correctly[13]. Chromosomes 1–5, 9 and 12 were closed, whereas chromosomes 6–8, 10, 11, 13 and 14 contained 3–37 gaps (most <2.5 kb) per chromosome at the beginning of genome annotation. Efforts to close the remaining gaps are continuing.

[1] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; [2] The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [3] Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA; [4] Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK; [5] University of Oxford, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK; [6] Department of Microbiology and Immunology, Drexel University College of Medicine, 2900 Queen Lane, Philadelphia, Pennsylvania 19129, USA; [7] School of Life Sciences, The Wellcome Trust Biocentre, The University of Dundee, Dundee DD1 5EH, UK; [8] Department of Biology and Genomics Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA; [9] Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Melbourne, VIC 3010, Australia; [10] Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA; [11] Department of Molecular and Cellular Biology, Berkeley Drosophila Genome Project, University of California, Berkeley, California 94720, USA; [12] The Center for the Advancement of Genomics, 1901 Research Boulevard, 6th Floor, Rockville, Maryland 20850, USA; [13] Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, Maryland 20910-7500, USA.

*Present addresses: Syngenta, Jealott's Hill International Research Centre, Bracknell, RG42 6EY, UK (S.B.); Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).

## Genome structure and content

The *P. falciparum* 3D7 nuclear genome is composed of 22.8 megabases (Mb) distributed among 14 chromosomes ranging in size from approximately 0.643 to 3.29 Mb (Fig. 1, and Supplementary Figs A–N). Thus the *P. falciparum* genome is almost twice the size of the genome of the fission yeast *Schizosaccharomyces pombe*. The overall (A + T) composition is 80.6%, and rises to ~90% in introns and intergenic regions. The structures of protein-encoding genes were predicted using several gene-finding programs and manually curated. Approximately 5,300 protein-encoding genes were identified, about the same as in *S. pombe* (Table 1, and Supplementary Table A). This suggests an average gene density in *P. falciparum* of 1 gene per 4,338 base pairs (bp), slightly higher than was found previously with chromosomes 2 and 3 (1 per 4,500 bp and 1 per 4,800 bp, respectively). The higher gene density reported here is probably the result of improved gene-finding software and larger training sets that enabled the detection of genes overlooked previously[8]. Introns were predicted in 54% of *P. falciparum* genes, a proportion roughly similar to that in *S. pombe* and *Dictyostelium discoideum*, but much higher than observed in *Saccharomyces cerevisiae* where only 5% of genes contain introns. Excluding introns, the mean length of *P. falciparum* genes was 2.3 kb, substantially larger than in the other organisms in which the average gene lengths range from 1.3 to 1.6 kb. *Plasmodium falciparum* genes showed a markedly greater proportion of genes (15.5%) longer than 4 kb compared to *S. pombe* and *S. cerevisiae* (3.0% and 3.6%, respectively). The explanation for the increased gene length in *P. falciparum* is not clear. Many of these large genes encode uncharacterized proteins that may be cytosolic proteins, as they do not possess recognizable signal peptides. No transposable elements or retrotransposons were identified.

Fifty-two per cent of the predicted gene products (2,731) were detected in cell lysates prepared from several stages of the parasite life cycle by high-resolution liquid chromatography and tandem mass spectrometry[14,15], including many predicted proteins with no similarity to proteins in other organisms. In addition, 49% of the genes overlapped (97% identity over at least 100 nucleotides) with expressed sequence tags (ESTs) derived from several life-cycle stages. As the proteomics and EST studies performed to date may not represent a complete sampling of all genes expressed during the complex life cycle of the parasite, this suggests that the annotation process identified substantial portions of most genes. However, in the absence of supporting EST or protein evidence, correct prediction of the 5′ ends of genes and genes with multiple small exons is challenging, and the gene models should be regarded as preliminary. Additional ESTs and full-length complementary DNA sequences[16] are required for the development of better training sets for gene-finding programs and the verification of the predicted genes.

The nuclear genome contains a full set of transfer RNA (tRNA) ligase genes, and 43 tRNAs were identified to bind all codons except TGT and TGC, coding for Cys; it is possible that these tRNAs are located within the currently unsequenced regions. All codons ending in C and T appear to be read by single tRNAs with a G in the first position, which is likely to read both codons via G:U wobble. Each anticodon occurs only once except for methionine (CAT), for which there are two copies, one for translation initiation and one for internal methionines, and the glycine (CCT) anticodon, which occurs twice. An unusual tRNA resembling a selenocysteinyl-tRNA was also found. A putative selenocysteine lyase was identified, which may provide selenium for synthesis of selenoproteins. Increased growth has been observed in selenium-supplemented *Plasmodium* culture[17].

In almost all other eukaryotic organisms sequenced to date, the tRNA genes exhibit extensive redundancy, the only exception being the intracellular parasite *Encephalitozoon cuniculi* which contains 44 tRNAs[18]. Often, the abundance of specific anticodons is correlated with the codon usage of the organism[19,20]. This is not the case in *P. falciparum*, which exhibits minimal redundancy of tRNAs. The mitochondrial genome of *Plasmodium* is small (about 6 kb) and encodes no tRNAs, so the mitochondrion must import tRNAs[21,22]. Through their import, cytoplasmic tRNAs may serve mitochondrial protein synthesis in a manner seen with other organisms[23,24]. The apicoplast genome appears to encode sufficient tRNAs for protein synthesis within the organelle[25].

Unlike many other eukaryotes, the malaria parasite genome does not contain long tandemly repeated arrays of ribosomal RNA (rRNA) genes. Instead, *Plasmodium* parasites contain several single 18S-5.8S-28S rRNA units distributed on different chromosomes.

Table 1 ***Plasmodium falciparum*** **nuclear genome summary and comparison to other organisms**

| Feature | Value | | | | |
|---|---|---|---|---|---|
| | *P. falciparum* | *S. pombe* | *S. cerevisiae* | *D. discoideum* | *A. thaliana* |
| Size (bp) | 22,853,764 | 12,462,637 | 12,495,682 | 8,100,000 | 115,409,949 |
| (G + C) content (%) | 19.4 | 36.0 | 38.3 | 22.2 | 34.9 |
| No. of genes | 5,268* | 4,929 | 5,770 | 2,799 | 25,498 |
| Mean gene length† (bp) | 2,283 | 1,426 | 1,424 | 1,626 | 1,310 |
| Gene density (bp per gene) | 4,338 | 2,528 | 2,088 | 2,600 | 4,526 |
| Per cent coding | 52.6 | 57.5 | 70.5 | 56.3 | 28.8 |
| Genes with introns (%) | 53.9 | 43 | 5.0 | 68 | 79 |
| Exons | | | | | |
| Number | 12,674 | ND | ND | 6,398 | 132,982 |
| No. per gene | 2.39 | ND | NA | 2.29 | 5.18 |
| (G + C) content (%) | 23.7 | 39.6 | 28.0 | 28.0 | ND |
| Mean length (bp) | 949 | ND | ND | 711 | 170 |
| Total length (bp) | 12,028,350 | ND | ND | 4,548,978 | 33,249,250 |
| Introns | | | | | |
| Number | 7,406 | 4,730 | 272 | 3,587 | 107,784 |
| (G + C) content (%) | 13.5 | ND | NA | 13.0 | ND |
| Mean length (bp) | 178.7 | 81 | NA | 177 | 170 |
| Total length (bp) | 1,323,509 | 383,130 | ND | 643,899 | 18,055,421 |
| Intergenic regions | | | | | |
| (G + C) content (%) | 13.6 | ND | ND | 14.0 | ND |
| Mean length (bp) | 1,694 | 952 | 515 | 786 | ND |
| RNAs | | | | | |
| No. of tRNA genes | 43 | 174 | ND | 73 | ND |
| No. of 5S rRNA genes | 3 | 30 | ND | NA | ND |
| No. of 5.8S, 18S and 28S rRNA units | 7 | 200–400 | ND | NA | 700–800 |

ND, not determined; NA, not applicable. 'No. of genes' for *D. discoideum* are for chromosome 2 (ref. 155) and in some cases represent extrapolations to the entire genome. Sources of data for the other organisms: *S. pombe*[65], *S. cerevisiae*[156], *D. discoideum*[155] and *A. thaliana*[157].
*70% of these genes matched expressed sequence tags or encoded proteins detected by proteomics analyses[14,15].
†Excluding introns.

The sequence encoded by a rRNA gene in one unit differs from the sequence of the corresponding rRNA in the other units. Furthermore, the expression of each rRNA unit is developmentally regulated, resulting in the expression of a different set of rRNAs at different stages of the parasite life cycle[26,27]. It is likely that by changing the properties of its ribosomes the parasite is able to alter the rate of translation, either globally or of specific messenger RNAs (mRNAs), thereby changing the rate of cell growth or altering patterns of cell development. The two types of rRNA genes previously described in *P. falciparum* are the S-type, expressed primarily in the mosquito vector, and the A-type, expressed primarily in the human host. Seven loci encoding rRNAs were identified in the genome sequence (Fig. 1). Two copies of the S-type rRNA genes are located on chromosomes 11 and 13, and two copies of the A-type genes are located on chromosomes 5 and 7. In addition, chromosome 1 contains a third, previously uncharacterized, rRNA unit that encodes 18S and 5.8S rRNAs that are almost identical to the S-type genes on chromosomes 11 and 13, but has a significantly divergent 28S rRNA gene (65% identity to the A-type and 75% identity to the S-type). The expression profiles of these genes are unknown. Chromosome 8 also contains two unusual rRNA gene units that contain 5.8S and 28S rRNA genes but do not encode 18S rRNAs; it is not known whether these genes are functional. The sequences of the 18S and 28S rRNA genes on chromosome 7 and the 28S rRNA gene on chromosome 8 are incomplete as they reside at contig ends. The 5S rRNA is encoded by three identical tandemly arrayed genes on chromosome 14.

## Chromosome structure

*Plasmodium falciparum* chromosomes vary considerably in length, with most of the variation occurring in the subtelomeric regions. Field isolates, even those from individuals residing in a single village[28], exhibit extensive size polymorphism that is thought to be due to recombination events between different parasite clones during meiosis in the mosquito[29]. Chromosome size variation is also observed in cultures of erythrocytic parasites, but is due to chromosome breakage and healing events and not to meiotic recombination[30,31]. Subtelomeric deletions often extend well into the chromosome, and in some cases alter the cell adhesion properties of the parasite owing to the loss of the gene(s) encoding adhesion molecules[32,33]. Because many genes involved in antigenic variation are located in the subtelomeric regions, an understanding of subtelomere structure and functional properties is essential for the elucidation of the mechanisms underlying the generation of antigenic diversity.

The subtelomeric regions of the chromosomes display a striking degree of conservation within the genome that is probably due to promiscuous inter-chromosomal exchange of subtelomeric regions. Subtelomeric exchanges occur in other eukaryotes[34–36], but the regions involved are much smaller (2.5–3.0 kb) in *S. cerevisiae* (data not shown). Previous studies of *P. falciparum* telomeres[37,38] suggested that they contained six blocks of repetitive sequences that were designated telomere-associated repetitive elements (TAREs 1–6).

Whole genome analysis reveals a larger (up to 120 kb), more complex, subtelomeric repeat structure than was observed previously. The conserved regions fall into five large subtelomeric blocks (SBs; Fig. 2). The sequences within blocks 2, 4 and 5 include many tandem repeats in addition to those described previously, as well as non-repetitive regions. Subtelomeric block 1 (SB-1, equivalent to TARE-1), contains the 7-bp telomeric repeat in a variable number of near-exact copies[39]. SB-2 contains several sub-blocks of repeats of different sizes, including TAREs 2–5 and other sequences. The beginning of SB-2 consists of about 1,000–1,300 bp of non-repetitive sequence, followed on some chromosomes by 2.5 copies of a 164-bp repeat. This is followed by another 300 bp of non-repetitive sequence, and then 10 copies of a 135-bp repeat, the main element of TARE-2. TARE-2 is followed by 200 bp of non-repetitive sequence, and then two copies of a highly conserved 63-bp repeat. SB-2 extends for another 6 kb that contains non-repetitive sequence as well as other tandem repeats. Only four of the 28 telomeres are missing SB-2, which always occurs immediately adjacent to SB-1. A notable feature of SB-2 is the conserved order and orientation of each repeat variant as well as the sequence homology extending throughout the block. For almost any two chromosomes that were examined, a consistently ordered series of unique, identical sequences of >30 bp that are distributed across SB-2 were identified, suggesting that SB-2 is a repeat with a complex internal structure occurring once per telomere.

SB-3 consists of the Rep20 element[40], a large block of highly variable copies of a 21-bp repeat. The tandem repeats in SB-3 occur in a random order (Fig. 2). SB-4 has not been described previously, although it does contain the previously described R-FA3 sequence[41]. SB-4 also includes a complex mix of short (<28-bp) tandem repeats, and a 105-bp repeat that occurs once in each subtelomere. Many telomeres contain one or more *var* (variant antigen) gene exons within this block, which appear as gaps in the alignment. In five subtelomeres, fragments of 2–4 kb from SB-4 are duplicated and inverted. SB-5 is found in half of the subtelomeres, does not contain tandem repeats, and extends up to 120 kb into some chromosomes. The arrangement and composition of the subtelomeric blocks suggests frequent recombination between the telomeres.

Centromeres have not been identified experimentally in malaria parasites. However, putative centromeres were identified by comparison of the sequences of chromosomes 2 and 3 (ref. 6). Eleven of the 14 chromosomes contained a single region of 2–3 kb with extremely high (A + T) content (>97%) and imperfect short tandem repeats, features resembling the regional *S. pombe* centromeres; the 3 chromosomes lacking such regions were incomplete.

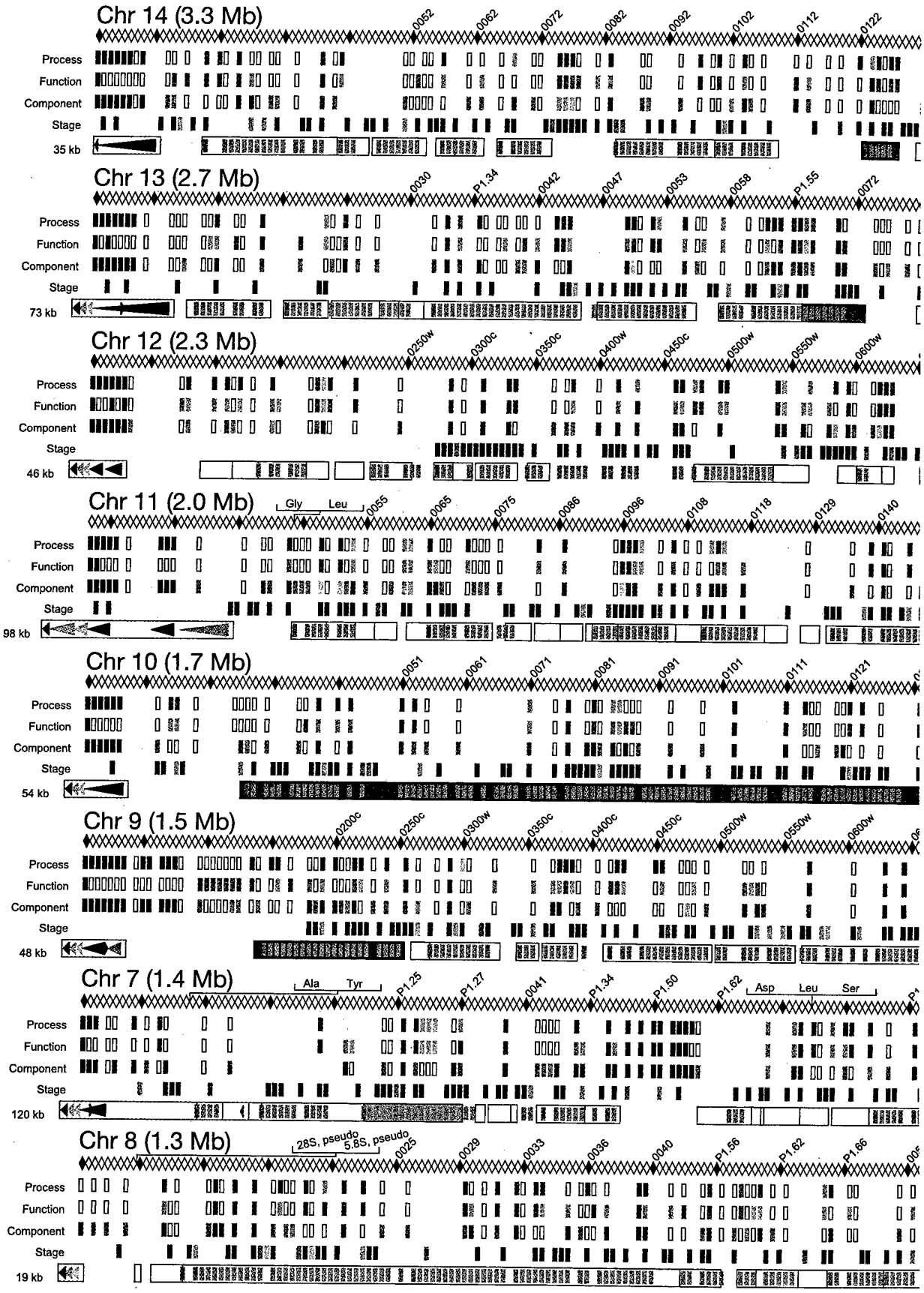
## The proteome

Of the 5,268 predicted proteins, about 60% (3,208 hypothetical proteins) did not have sufficient similarity to proteins in other organisms to justify provision of functional assignments (Table 2). This is similar to what was found previously with chromosomes 2 and 3 (refs 5, 6). Thus, almost two-thirds of the proteins appear to be unique to this organism, a proportion much higher than observed in other eukaryotes. This may be a reflection of the greater evolutionary distance between *Plasmodium* and other eukaryotes that have been sequenced, exacerbated by the reduction of sequence similarity due to the (A + T) richness of the genome. Another 257 proteins (5%) had significant similarity to hypothetical proteins in other organisms. Thirty-one per cent (1,631) of the predicted proteins had one or more transmembrane domains, and 17.3% (911) of the proteins possessed putative signal peptides or signal anchors.

The Gene Ontology (GO)[42] database is a controlled vocabulary that describes the roles of genes and gene products in organisms. GO terms were assigned manually to 2,134 gene products (40%)

**Figure 1** Schematic representation of the *P. falciparum* 3D7 genome. Protein-encoding genes are indicated by open diamonds. All genes are depicted at the same scale regardless of their size or structure. The labels indicate the name for each gene. The rows of coloured rectangles represent, from top to bottom for each chromosome, the high-level Gene Ontology assignment for each gene in the 'biological process', 'molecular function', and 'cellular component' ontologies[42]; the life-cycle stage(s) at which each predicted gene product has been detected by proteomics techniques[14,15]; and *Plasmodium yoelii yoelii* genes that exhibit conserved sequence and organization with genes in *P. falciparum*, as shown by a position effect analysis. Rectangles surrounding clusters of *P. yoelii* genes indicate genes shown to be linked in the *P. y. yoelii* genome[165]. Boxes containing coloured arrowheads at the ends of each chromosome indicate subtelomeric blocks (SBs; see text and Fig. 2).

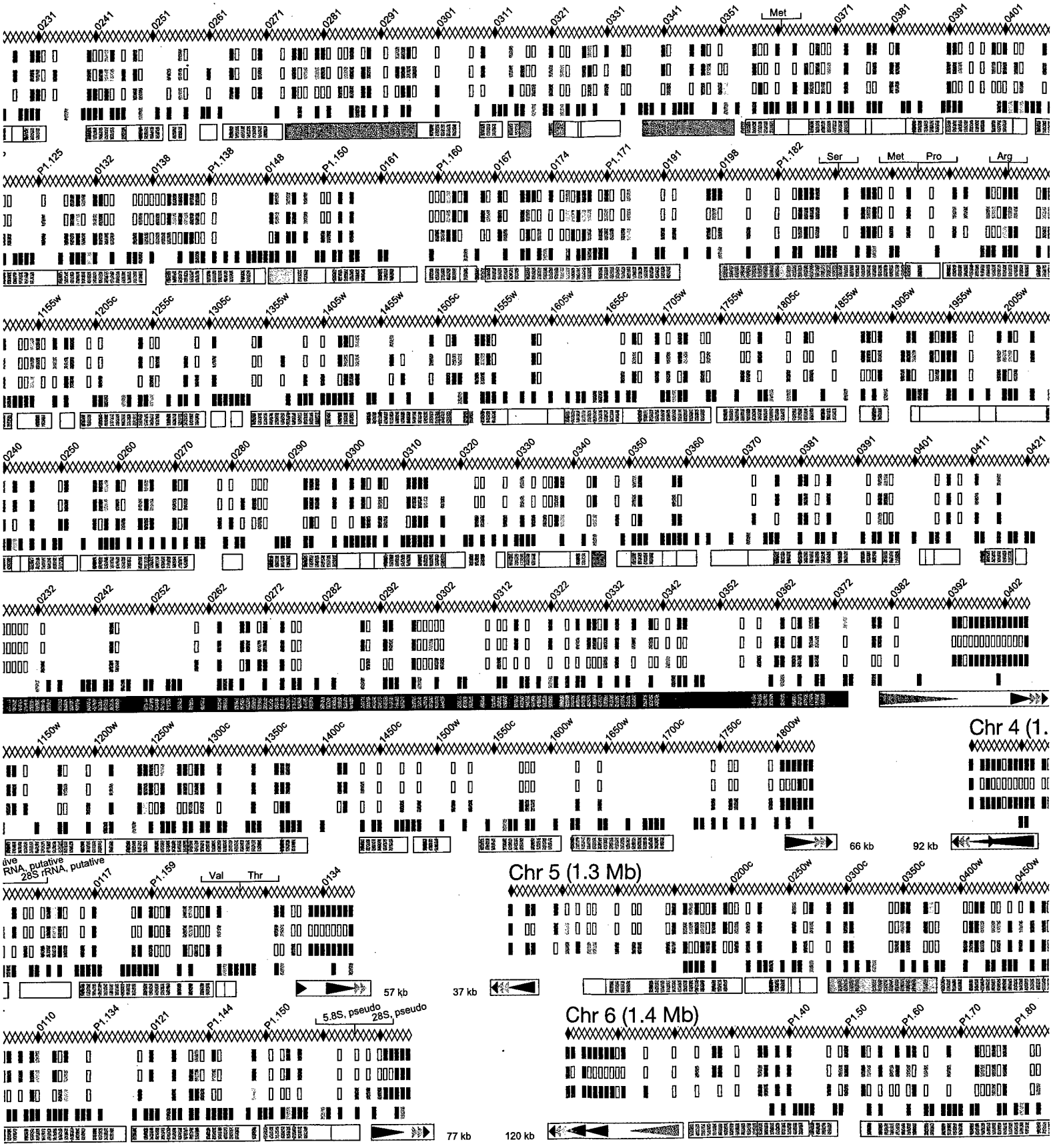
Figure 1

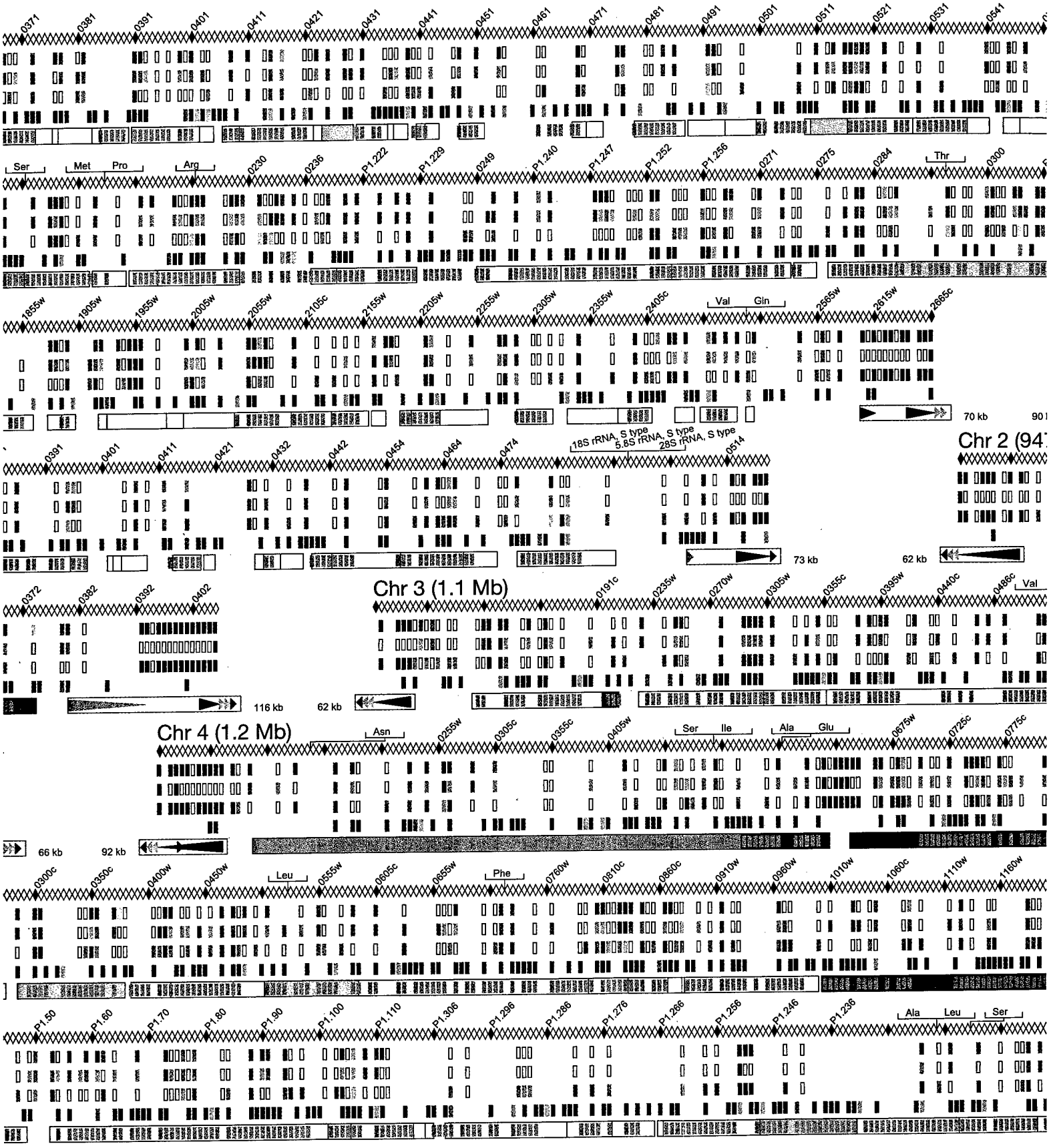

Chr 14 (3.3 Mb)
0052
0062
0072
0082
0092
0102
0112
0122
Process
Function
Component
Stage
35 kb
Chr 13 (2.7 Mb)
0030
P1.34
0042
0047
0053
0058
P1.55
0072
Process
Function
Component
Stage
73 kb
Chr 12 (2.3 Mb)
0250w
0300c
0350c
0400w
0450c
0500w
0550w
0600w
Process
Function
Component
Stage
46 kb
Chr 11 (2.0 Mb)
Gly
Leu
0055
0065
0075
0086
0096
0108
0118
0129
0140
Process
Function
Component
Stage
98 kb
Chr 10 (1.7 Mb)
0051
0061
0071
0081
0091
0101
0111
0121
Process
Function
Component
Stage
54 kb
Chr 9 (1.5 Mb)
0200c
0250c
0300w
0350c
0400c
0450c
0500w
0550w
0600w
Process
Function
Component
Stage
48 kb
Chr 7 (1.4 Mb)
Ala
Tyr
P1.25
P1.27
0041
P1.34
P1.50
P1.62
Asp
Leu
Ser
Process
Function
Component
Stage
120 kb
Chr 8 (1.3 Mb)
28S, pseudo
5.8S, pseudo
0025
0029
0033
0036
0040
P1.56
P1.82
P1.86
Process
Function
Component
Stage
19 kb

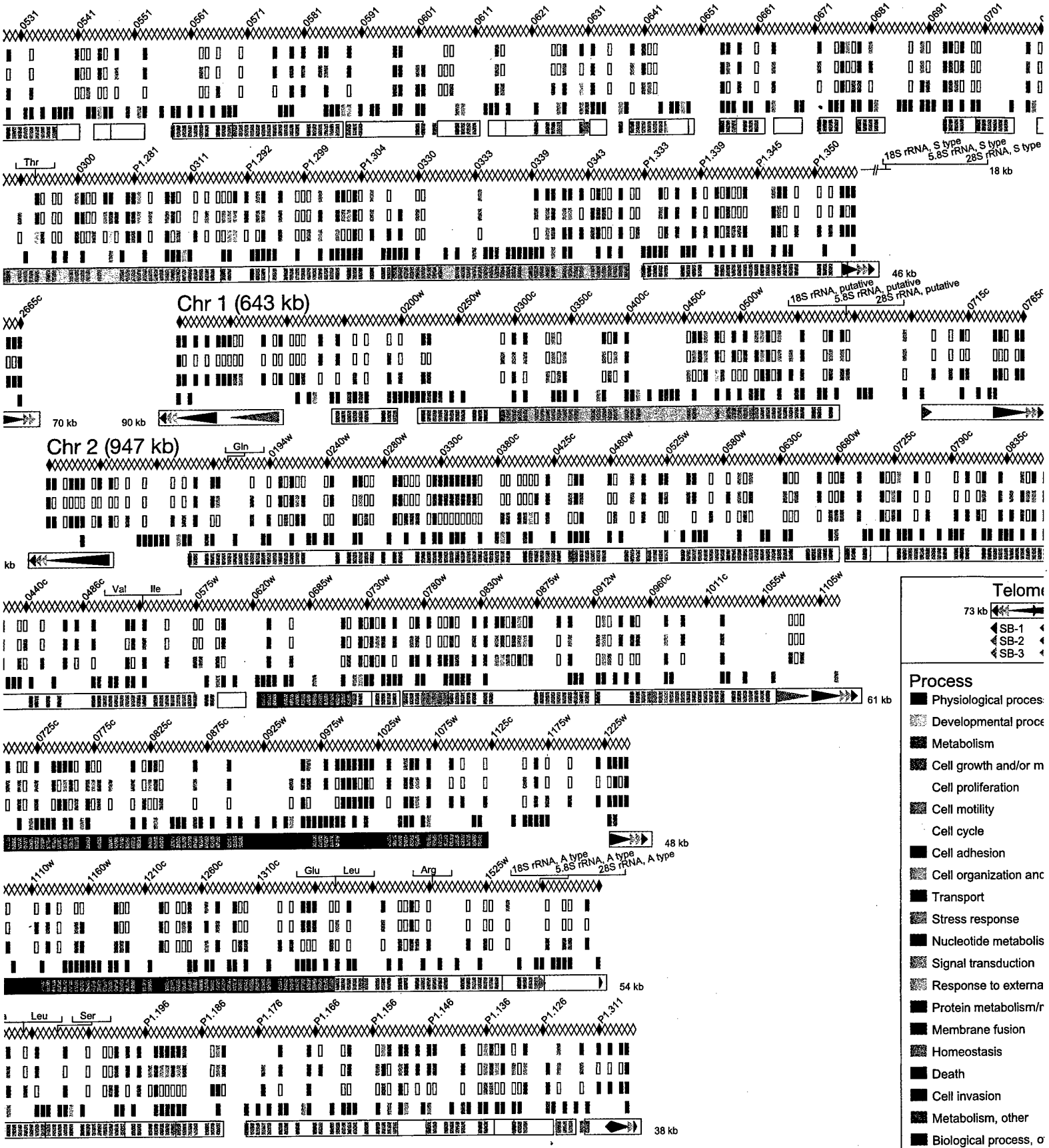
gure 1


Pro
5S rRNA
5S rRNA
5S rRNA
Pro
Asp
Leu
Ser
18S rRNA, putative
5.8S rRNA, putative
28S rRNA, putative

(2)


Met
Ser
Met
Pro
Arg
Chr 4 (1.
66 kb
92 kb
Val
Thr
Chr 5 (1.3 Mb)
57 kb
37 kb
5.8S, pseudo
28S, pseudo
Chr 6 (1.4 Mb)
77 kb
120 kb


Ser
Met
Pro
Arg
Thr
Val
Gln
Chr 2 (94
70 kb
73 kb
62 kb
18S rRNA, S type
5.8S rRNA, S type
28S rRNA, S type
Chr 3 (1.1 Mb)
116 kb
62 kb
Val
Chr 4 (1.2 Mb)
Asn
Ser
Ile
Ala
Glu
66 kb
92 kb
Leu
Phe
Ala
Leu
Ser


Chr 1 (643 kb)
Chr 2 (947 kb)
Thr
Gln
Val
Ile
Glu
Leu
Arg
Leu
Ser
18S rRNA, S type
5.8S rRNA, S type
28S rRNA, S type
18S rRNA, putative
5.8S rRNA, putative
28S rRNA, putative
18S rRNA, A type
5.8S rRNA, A type
28S rRNA, A type
18 kb
46 kb
70 kb
90 kb
61 kb
48 kb
54 kb
38 kb
Telom
73 kb
SB-1
SB-2
SB-3
Process
Physiological proces
Developmental proce
Metabolism
Cell growth and/or m
Cell proliferation
Cell motility
Cell cycle
Cell adhesion
Cell organization anc
Transport
Stress response
Nucleotide metabolis
Signal transduction
Response to externa
Protein metabolism/r
Membrane fusion
Homeostasis
Death
Cell invasion
Metabolism, other
Biological process, o
Cell communications
Biological process ur


0611
0621
0631
0641
0651
0661
0671
0681
0691
0701
0711
0721
0731
0741
0752
0762
0772
97 kb
0333
0339
0343
P1.333
P1.339
P1.345
P1.350
18S rRNA, S type
5.8S rRNA, S type
28S rRNA, S type
18 kb
46 kb
0300c
0350c
0400c
0450c
0500w
18S rRNA, putative
5.8S rRNA, putative
28S rRNA, putative
0715c
0765c
96 kb
0380c
0425c
0480w
0525w
0580w
0630c
0680w
0725c
0790c
0835c
0886c
0923c
0960c
0990c
1040w
73 kb
0830w
0875w
0912w
0960c
1011c
1055w
1105w
61 kb
1125c
1175w
1225w
48 kb
1525w
18S rRNA, A type
5.8S rRNA, A type
28S rRNA, A type
54 kb
P1.136
P1.126
P1.311
38 kb
Telomeres
73 kb
Telomere region
SB-1
SB-2
SB-3
SB-4
SB-4F
SB-5
rRNA and tRNA genes
18S rRNA, A type
5.8S rRNA, A type
28S rRNA, A type
Val
Thr
Process
Physiological processes
Developmental processes
Metabolism
Cell growth and/or maintenance
Cell proliferation
Cell motility
Cell cycle
Cell adhesion
Cell organization and biogenesis
Transport
Stress response
Nucleotide metabolism
Signal transduction
Response to external stimulus
Protein metabolism/modification
Membrane fusion
Homeostasis
Death
Cell invasion
Metabolism, other
Biological process, other
Cell communications, other
Biological process unknown
Function
Enzyme
Enzyme regulator
Defence/immunity protein
Structural molecule
Cell adhesion molecule
Chaperone
Transporter
Ligand binding or carrier
Transcription regulator
Translation regulator
Molecular function, other
Molecular function unknown
Component
Nucleus
Cytoplasm
Extracellular
Mitochondrion
Plastid
Membrane
Vacuole
Cytoplasm, other
Intracellular, other
Cell, other
Unlocalized
Cellular component unknown
P. yoelii synteny
Syntenic group
Syntenic gene
Chr 2
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Not determined
Stage
Gametocyte
Sporozoite
Gametocyte + sporozoite
Erythrocytic
Erythrocytic + sporozoite
Erythrocytic + gametocyte
All

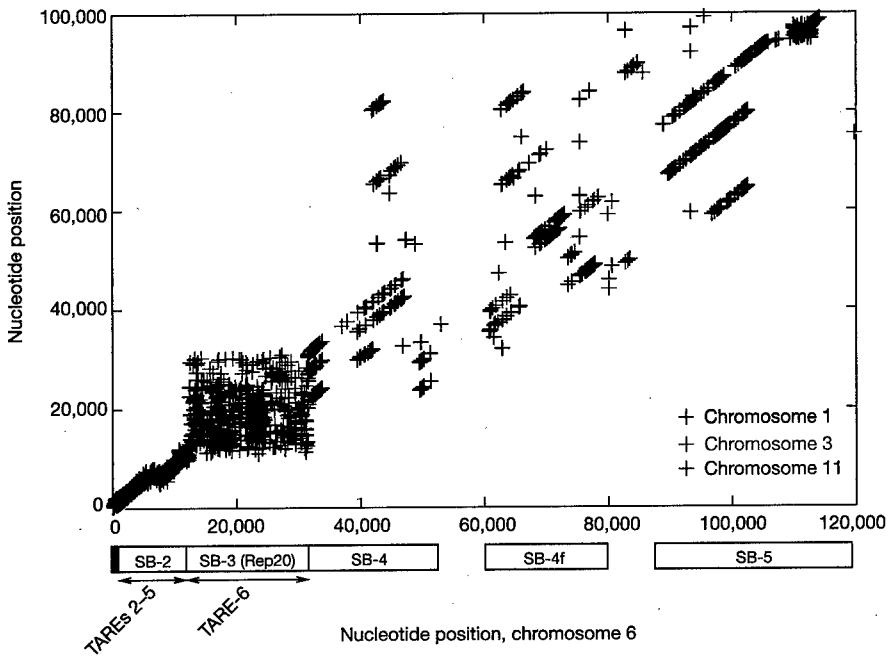



**Figure 2** Alignment of subtelomeric regions of chromosomes 1, 3, 6 and 11. MUMmer2[152] alignments showing exact matches between the left subtelomeric regions of chromosome 6 (horizontal axis) and chromosomes 11 (red), 1 (blue) and 3 (green), illustrating the conserved synteny between all telomeres. Each point represents an exact match of 40 bp or longer that is shared by two chromosomes and is not found anywhere else on either chromosome. Each collinear series of points along a diagonal represents an aligned region. SB, subtelomeric block; TARE, telomere-associated repetitive element.

and a comparison of annotation with high-level GO terms for both *S. cerevisiae* and *P. falciparum* is shown in Fig. 3. In almost all categories, higher values can be seen for *S. cerevisiae*, reflecting the greater proportion of the genome that has been characterized compared to *P. falciparum*. There are two exceptions to this pattern that reflect processes specifically connected with the parasite life cycle. At least 1.3% of *P. falciparum* genes are involved in cell-to-cell adhesion or the invasion of host cells. As discussed below (see 'Immune evasion'), *P. falciparum* has 208 genes (3.9%) known to be involved in the evasion of the host immune system. This is reflected in the assignment of many more gene products to the GO term 'physiological processes' in *P. falciparum* than in *S. cerevisiae* (Fig. 3). The comparison with *S. cerevisiae* also reveals that particular categories in *P. falciparum* appear to be under-represented. Sporulation and cell budding are obvious examples (they are included in the category 'other cell growth and/or maintenance'), but very few genes in *P. falciparum* were associated with the 'cell organization and biogenesis', the 'cell cycle', or 'transcription factor' categories compared to *S. cerevisiae* (Fig. 3). These differences do not necessarily imply that fewer malaria genes are involved in these processes, but highlight areas of malaria biology where knowledge is limited.

Table 2 **The *P. falciparum* proteome**

| Feature | Number | Per cent |
|---|---|---|
| Total predicted proteins | 5,268 | |
| Hypothetical proteins | 3,208 | 60.9 |
| InterPro matches | 2,650 | 52.8 |
| Pfam matches | 1,746 | 33.1 |
| Gene Ontology | | |
| Process | 1,301 | 24.7 |
| Function | 1,244 | 23.6 |
| Component | 2,412 | 45.8 |
| Targeted to apicoplast | 551 | 10.4 |
| Targeted to mitochondrion | 246 | 4.7 |
| Structural features | | |
| Transmembrane domain(s) | 1,631 | 31.0 |
| Signal peptide | 544 | 10.3 |
| Signal anchor | 367 | 7.0 |
| Non-secretory protein | 4,357 | 82.7 |

Of the apicoplast-targeted proteins, 126 were judged on the basis of experimental evidence or the predictions of multiple programs[61,158] to be localized to the apicoplast with high confidence. Predicted apicoplast localization for 425 other proteins is based on an analysis using only one method and is of lower confidence. Predicted mitochondrial localization was based upon BLASTP searches of *S. cerevisiae* mitochondrion-targeted proteins[159] and TargetP[158] and MitoProtII[160] predictions; 148 genes were judged to be targeted to the mitochondrion with a high or medium confidence level, and an additional 98 genes with a lower confidence of mitochondrial targeting. Other specialized searches used the following programs and databases: InterPro[161]; Pfam[162]; Gene Ontology[42]; transmembrane domains, TMHMM[163]; signal peptides and signal anchors, SignalP-2.0[164].

## The apicoplast

Malaria parasites and other members of the phylum apicomplexa harbour a relict plastid, homologous to the chloroplasts of plants and algae[25,43,44]. The 'apicoplast' is essential for parasite survival[45,46], but its exact role is unclear. The apicoplast is known to function in the anabolic synthesis of fatty acids[5,47,48], isoprenoids[49] and haeme[50,51], suggesting that one or more of these compounds could be exported from the apicoplast, as is known to occur in plant plastids. The apicoplast arose through a process of secondary endosymbiosis[52–55], in which the ancestor of all apicomplexan parasites engulfed a eukaryotic alga, and retained the algal plastid, itself the product of a prior endosymbiotic event[56]. The 35-kb apicoplast genome encodes only 30 proteins[25], but as in mitochondria and chloroplasts, the apicoplast proteome is supplemented by proteins encoded in the nuclear genome and post-translationally targeted into the organelle by the use of a bipartite targeting signal, consisting of an amino-terminal secretory signal sequence, followed by a plastid transit peptide[55,57–60].

In total, 551 nuclear-encoded proteins (~10% of the predicted nuclear encoded proteins) that may be targeted to the apicoplast were identified using bioinformatic[61] and laboratory-based methods. Apicoplast targeting of a few proteins has been verified by antibody localization and by the targeting of fluorescent fusion proteins to the apicoplast in transgenic *P. falciparum* or *Toxoplasma gondii*[47] parasites. Some proteins may be targeted to both the apicoplast and mitochondrion, as suggested by the observation that the total number of tRNA ligases is inadequate for independent

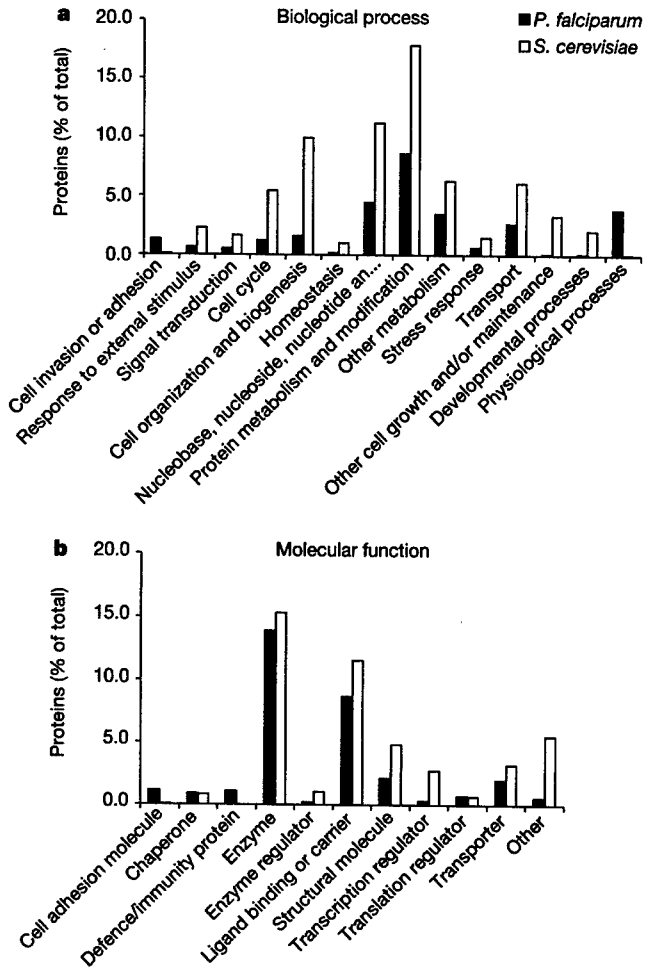



**Figure 3** Gene Ontology classifications. Classification of *P. falciparum* proteins according to the 'biological process' (**a**) and 'molecular function' (**b**) ontologies of the Gene Ontology system[42].

protein synthesis in the cytoplasm, mitochondrion and apicoplast. In plants, some proteins lack a transit peptide but are targeted to plastids via an unknown process. Proteins that use an alternative targeting pathway in *P. falciparum* would have escaped detection with the methods used.

Nuclear-encoded apicoplast proteins include housekeeping enzymes involved in DNA replication and repair, transcription, translation and post-translational modifications, cofactor synthesis, protein import, protein turnover, and specific metabolic and transport activities. No genes for photosynthesis or light perception are apparent, although ferredoxin and ferredoxin-NADP reductase are present as vestiges of photosystem I, and probably serve to recycle reducing equivalents[62]. About 60% of the putative apicoplast-targeted proteins are of unknown function. Several metabolic pathways in the organelle are distinct from host pathways and offer potential parasite-specific targets for drug therapy[63] (see 'Metabolism' and 'Transport' sections).

## Evolution

Comparative genome analysis with other eukaryotes for which the complete genome is available (excluding the parasite *E. cuniculi*) revealed that, in terms of overall genome content, *P. falciparum* is slightly more similar to *Arabidopsis thaliana* than to other taxa. Although this is consistent with phylogenetic studies[64], it could also be due to the presence in the *P. falciparum* nuclear genome of genes derived from plastids or from the nuclear genome of the secondary endosymbiont. Thus the apparent affinity of *Plasmodium* and *Arabidopsis* might not reflect the true phylogenetic history of the *P. falciparum* lineage. Comparative genomic analysis was also used to identify genes apparently duplicated in the *P. falciparum* lineage since it split from the lineages represented by the other completed genomes (Supplementary Table B).

There are 237 *P. falciparum* proteins with strong matches to proteins in all completed eukaryotic genomes but no matches to proteins, even at low stringency, in any complete prokaryotic proteome (Supplementary Table C). These proteins help to define the differences between eukaryotes and prokaryotes. Proteins in this list include those with roles in cytoskeleton construction and maintenance, chromatin packaging and modification, cell cycle regulation, intracellular signalling, transcription, translation, replication, and many proteins of unknown function. This list overlaps with, but is somewhat larger than, the list generated by an analysis of the *S. pombe* genome[65]. The differences are probably due in part to the different stringencies used to identify the presence or absence of homologues in the two studies.

A large number of nuclear-encoded genes in most eukaryotic species trace their evolutionary origins to genes from organelles that have been transferred to the nucleus during the course of eukaryotic evolution. Similarity searches against other complete genomes were used to identify *P. falciparum* nuclear-encoded genes that may be derived from organellar genomes. Because similarity searches are not an ideal method for inferring evolutionary relatedness[66], phylogenetic analysis was used to gain a more accurate picture of the evolutionary history of these genes. Out of 200 candidates examined, 60 genes were identified as being of probable mitochondrial origin. The proteins encoded by these genes include many with known or expected mitochondrial functions (for example, the tricarboxylic acid (TCA) cycle, protein translation, oxidative damage protection, the synthesis of haem, ubiquinone and pyrimidines), as well as proteins of unknown function. Out of 300 candidates examined, 30 were identified as being of probable plastid origin, including genes with predicted roles in transcription and translation, protein cleavage and degradation, the synthesis of isoprenoids and fatty acids, and those encoding four subunits of the pyruvate dehydrogenase complex. The origin of many candidate organelle-derived genes could not be conclusively determined, in part due to the problems inherent in analysing genes of very high (A + T) content. Nevertheless, it appears likely that the total number of plastid-derived genes in *P. falciparum* will be significantly lower than that in the plant *A. thaliana* (estimated to be over 1,000). Phylogenetic analysis reveals that, as with the *A. thaliana* plastid, many of the genes predicted to be targeted to the apicoplast are apparently not of plastid origin. Of 333 putative apicoplast-targeted genes for which trees were constructed, only 26 could be assigned a probable plastid origin. In contrast, 35 were assigned a probable mitochondrial origin and another 85 might be of mitochondrial origin but are probably not of plastid origin (they group with eukaryotes that have not had plastids in their history, such as humans and fungi, but the relationship to mitochondrial ancestors is not clear). The apparent non-plastid origin of these genes could either be due to inaccuracies in the targeting predictions or to the co-option of genes derived from the mitochondria or the nucleus to function in the plastid, as has been shown to occur in some plant species[67].

## Metabolism

Biochemical studies of the malaria parasite have been restricted primarily to the intra-erythrocytic stage of the life cycle, owing to the difficulty of obtaining suitable quantities of material from the other life-cycle stages. Analysis of the genome sequence provides a global view of the metabolic potential of *P. falciparum* irrespective of the life-cycle stage (Fig. 4). Of the 5,268 predicted proteins, 733 (~14%) were identified as enzymes, of which 435 (~8%) were assigned Enzyme Commission (EC) numbers. This is considerably

fewer than the roughly one-quarter to one-third of the genes in bacterial and archaeal genomes that can be mapped to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway diagrams[68], or the 17% of *S. cerevisiae* open reading frames that can be assigned EC numbers. This suggests either that *P. falciparum* has a smaller proportion of its genome devoted to enzymes, or that enzymes are more difficult to identify in *P. falciparum* by sequence similarity methods. (This difficulty can be attributed either to the great evolutionary distance between *P. falciparum* and other well-studied organisms, or to the high (A + T) content of the genome.) A few genes might have escaped detection because they were located in the small regions of the genome that remain to be sequenced (Table 1). However, many biochemical pathways could be reconstructed in their entirety, suggesting that the similarity-searching approach was for the most part successful, and that the relative paucity of enzymes in *P. falciparum* may be related to its parasitic life-style. A similar picture has emerged in the analysis of transporters (see 'Transport').

In erythrocytic stages, *P. falciparum* relies principally on anaerobic glycolysis for energy production, with regeneration of $NAD^+$ by conversion of pyruvate to lactate[69]. Genes encoding all of the enzymes necessary for a functional glycolytic pathway were identified, including a phosphofructokinase (PFK) that has sequence similarity to the pyrophosphate-dependent class of enzymes but which is probably ATP-dependent on the basis of the characterization of the homologous enzyme in *Plasmodium berghei*[70,71]. A second putative pyrophosphate-dependent PFK was also identified which possessed N- and carboxy-terminal extensions that could represent targeting sequences.

A gene encoding fructose bisphosphatase could not be detected, suggesting that gluconeogenesis is absent, as are enzymes for synthesis of trehalose, glycogen or other carbohydrate stores. Candidate genes for all but one enzyme of the conventional pentose

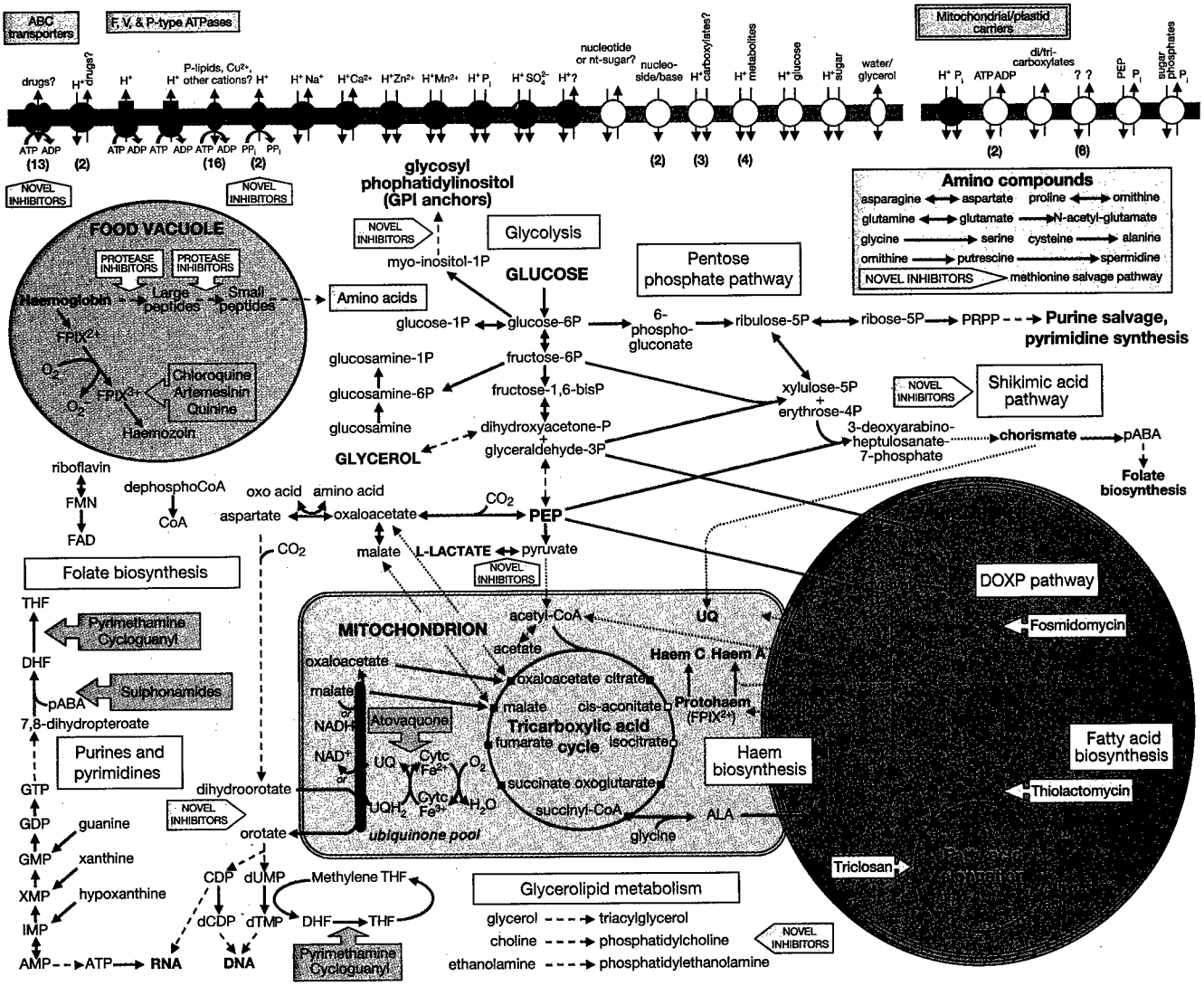



**Figure 4** Overview of metabolism and transport in *P. falciparum*. Glucose and glycerol provide the major carbon sources for malaria parasites. Metabolic steps are indicated by arrows, with broken lines indicating multiple intervening steps not shown; dotted arrows indicate incomplete, unknown or questionable pathways. Known or potential organellar localization is shown for pathways associated with the food vacuole, mitochondrion and apicoplast. Small white squares indicate TCA (tricarboxylic acid) cycle metabolites that may be derived from outside the mitochondrion. Fuschia block arrows indicate the steps inhibited by antimalarials; grey block arrows highlight potential drug targets. Transporters are grouped by substrate specificity: inorganic cations (green), inorganic anions (magenta), organic nutrients (yellow), drug efflux and other (black). Arrows indicate direction of transport for substrates (and coupling ions, where appropriate). Numbers in parentheses indicate the presence of multiple transporter genes with similar substrate predictions. Membrane transporters of unknown or putative subcellular localization are shown in a generic membrane (blue bar). Abbreviations: ACP, acyl carrier protein; ALA, aminolevulinic acid; CoA, coenzyme A; DHF, dihydrofolate; DOXP, deoxyxylulose phosphate; $FPIX^{2+}$ and $FPIX^{3+}$, ferro- and ferriprotoporphyrin IX, respectively; pABA, *para*-aminobenzoic acid; PEP, phosphoenolpyruvate; $P_i$, phosphate; $PP_i$, pyrophosphate; PRPP, phosphoribosyl pyrophosphate; THF, tetrahydrofolate; UQ, ubiquinone.

phosphate pathway were found. These include a bifunctional glucose-6-phosphate dehydrogenase/6-phosphogluconate dehydrogenase required to generate NADPH and ribose 5-phosphate for other biosynthetic pathways[72,73]. Transaldolase appears to be absent, but erythrose 4-phosphate required for the chorismate pathway could probably be generated from the glycolytic intermediates fructose 6-phosphate and glyceraldehyde 3-phosphate via a putative transketolase (Fig. 4).

The genes necessary for a complete TCA cycle, including a complete pyruvate dehydrogenase complex, were identified. However, it remains unclear whether the TCA cycle is used for the full oxidation of products of glycolysis, or whether it is used to supply intermediates for other biosynthetic pathways. The pyruvate dehydrogenase complex seems to be localized in the apicoplast, and the only protein with significant similarity to aconitases has been reported to be a cytosolic iron-response element binding protein that did not possess aconitase activity[74]. Also, malate dehydrogenase appears to be cytosolic rather than mitochondrial, even though it seems to have originated from the mitochondrial genome[75]. Genes encoding malate-quinone oxidoreductase and type I fumarate dehydratase are present. Malate-quinone oxidoreductase, which is probably targeted to the mitochondrion, may well replace malate dehydrogenase in the TCA cycle, as it does in *Helicobacter pylori*. A gene encoding phosphoenolpyruvate carboxylase (PEPC) was also found. Like bacteria and plants, *P. falciparum* may cope with a drain of TCA cycle intermediates by using phosphoenolpyruvate (PEP) to replenish oxaloacetate (Fig. 4). This would seem to be supported by reports of $CO_2$-incorporating activity in asexual stage parasite cultures[76]. Thus, the TCA cycle appears to be unconventional in erythrocytic stages, and may serve mainly to synthesize succinyl-CoA, which in turn can be used in the haem biosynthesis pathway.

Genes encoding all subunits of the catalytic $F_1$ portion of ATP synthase, the protein that confers oligomycin sensitivity, and the gene that encodes the proteolipid subunit *c* for the $F_0$ portion of ATP synthase, were detected in the parasite genome. The $F_0$ *a* and *b* subunits could not be detected, raising the question as to whether the ATP synthase is functional. Because parts of the genome sequence are incomplete, the presence of the *a* and *b* subunits could not be ruled out. Erythrocytic parasites derive ATP through glycolysis and the mitochondrial contribution to the ATP pool in these stages appears to be minimal[77,78]. It is possible that the ATP synthase functions in the insect or sexual stages of the parasite. However, in the absence of the $F_0$ *a* and *b* subunits, an ATP synthase cannot use the proton gradient[79].

A functional mitochondrion requires the generation of an electrochemical gradient across the inner membrane. But the *P. falciparum* genome seems to lack genes encoding components of a conventional NADH dehydrogenase complex I. Instead, a single subunit NADH dehydrogenase gene specifies an enzyme that can accomplish ubiquinone reduction without proton pumping, thus constituting a non-electrogenic step. Other dehydrogenases targeted to the mitochondrion also serve to reduce ubiquinone in *P. falciparum*, including dihydroorotate dehydrogenase, a critical enzyme in the essential pyrimidine biosynthesis pathway[80]. The parasite genome contains some genes specifying ubiquinone synthesis enzymes, in agreement with recent metabolic labelling studies[81]. Re-oxidation of ubiquinol is carried out by the cytochrome *bc1* complex that transfers electrons to cytochrome *c*, and is accompanied by proton translocation[82]. Apocytochrome *b* of this complex is encoded by the mitochondrial genome[21,22], but the rest of the components are encoded by nuclear genes. Ubiquinol cycling is a critical step in mitochondrial physiology, and its selective inhibition by hydroxynaphthoquinones is the basis for their antimalarial action[83]. The final step in electron transport is carried out by the proton-pumping cytochrome *c* oxidase complex, of which only two subunits are encoded in the mitochondrial DNA (mtDNA). In most eukaryotes, subunit II of cytochrome *c* oxidase is encoded by a gene on the mitochondrial genome. In *P. falciparum*, however, the *coxII* gene is divided such that the N-terminal portion is encoded on chromosome 13 and the C-terminal portion on chromosome 14. A similar division of the *coxII* gene is also seen in the unicellular alga, *Chlamydomonas reinhardtii*[84]. An alternative oxidase that transfers electrons directly from ubiquinol to oxygen has been seen in plants as well in many protists, and an earlier biochemical study suggested its presence in *P. falciparum*[85]. The genome sequence, however, fails to reveal such an oxidase gene.

Biochemical, genetic and chemotherapeutic data suggest that malaria and other apicomplexan parasites synthesize chorismate from erythrose 4-phosphate and phosphoenolpyruvate via the shikimate pathway[86–89]. It was initially suggested that the pathway was located in the apicoplast[88], but chorismate synthase is phylogenetically unrelated to plastid isoforms[90] and has subsequently been localized to the cytosol[91]. The genes for the preceding enzymes in the pathway could not be identified with certainty, but a BLASTP search with the *S. cerevisiae* arom polypeptide[92], which catalyses 5 of the preceding steps, identified a protein with a low level of similarity (E value $7.9 \times 10^{-8}$).

In many organisms, chorismate is the pivotal precursor to several pathways, including the biosynthesis of aromatic amino acids and ubiquinone. We found no evidence, on the basis of similarity searches, for a role of chorismate in the synthesis of tryptophan, tyrosine or phenylalanine, although *para*-aminobenzoate (pABA) synthase does have a high degree of similarity to anthranilate (2-amino benzoate) synthase, the enzyme catalysing the first step in tryptophan synthesis from chorismate. In accordance with the supposition that the malaria parasite obtains all of its amino acids either by salvage from the host or by globin digestion, we found no enzymes required for the synthesis of other amino acids with the exception of enzymes required for glycine–serine, cysteine–alanine, aspartate–asparagine, proline–ornithine and glutamine–glutamate interconversions. In addition to pABA synthase, all but one of the enzymes (dihydroneopterin aldolase) required for *de novo* synthesis of folate from GTP were identified.

Several studies have shown that the erythrocytic stages of *P. falciparum* are incapable of *de novo* purine synthesis (reviewed in ref. 80). This statement can now be extended to all life-cycle stages, as only adenylsuccinate lyase, one of the 10 enzymes required to make inosine monophosphate (IMP) from phosphoribosyl pyrophosphate, was identified. This enzyme also plays a role in purine salvage by converting IMP to AMP. Purine transporters and enzymes for the interconversion of purine bases and nucleosides are also present. The parasite can synthesize pyrimidines *de novo* from glutamine, bicarbonate and aspartate, and the genes for each step are present. Deoxyribonucleotides are formed via an aerobic ribonucleoside diphosphate reductase[93,94], which is linked via thioredoxin to thioredoxin reductase. Gene knockout experiments have recently shown that thioredoxin reductase is essential for parasite survival[95].

The intraerythrocytic stages of the malaria parasite uses haemoglobin from the erythrocyte cytoplasm as a food source, hydrolysing globin to small peptides, and releasing haem that is detoxified in the form of haemazoin. Although large amounts of haem are toxic to the parasite, *de novo* haem biosynthesis has been reported[96] and presumably provides a mechanism by which the parasite can segregate host-derived haem from haem required for synthesis of its own iron-containing proteins. However, it has been unclear whether *de novo* synthesis occurs using imported host enzymes[97] or parasite-derived enzymes. Genes encoding the first two enzymes in the haem biosynthetic pathway, aminolevulinate synthase[98] and aminolevulinate dehydratase[99], were cloned previously, and genes encoding every other enzyme in the pathway except for uroporphyrinogen-III synthase were found (Fig. 4).

Haem and iron–sulphur clusters form redox prosthetic groups for a wide range of proteins, many of which are localized to the

mitochondrion and apicoplast. The parasite genome appears to encode enzymes required for the synthesis of these molecules. There are two putative cysteine desulphurase genes, one which also has homology to selenocysteine lyase and may be targeted to the mitochondrion, and the second which may be targeted to the apicoplast, suggesting organelle specific generation of elemental sulphur to be used in Fe–S cluster proteins. The subcellular localization of the enzymes involved in haem synthesis is uncertain. Ferrochelatase and two haem lyases are likely to be localized in the mitochondrion.

The role of the apicoplast in type II fatty-acid biosynthesis was described previously[5,47]. The genes encoding all enzymes in the pathway have now been elucidated, except for a thioesterase required for chain termination. No evidence was found for the associative (type I) pathway for fatty-acid biosynthesis common to most eukaryotes. The apicoplast also houses the machinery for mevalonate-independent isoprenoid synthesis. Because it is not present in mammals, the biosynthesis of isopentyl diphosphate from pyruvate and glyceraldehyde-3-phosphate provides several attractive targets for chemotherapy. Three enzymes in the pathway have been identified, including 1-deoxy-D-xylulose-5-phosphate synthase, 1-deoxy-D-xylulose-5-phosphate reductoisomerase[49], and 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase[100,101]. One predicted protein was similar to the fourth enzyme, 2C-methyl-D-erythritol-4-phosphate cytidyltransferase (BLASTP E value $9.6 \times 10^{-15}$).

## Transport

On the basis of genome analysis, *P. falciparum* possesses a very limited repertoire of membrane transporters, particularly for uptake of organic nutrients, compared to other sequenced eukaryotes (Fig. 5). For instance, there are only six *P. falciparum* members of the major facilitator superfamily (MFS) and one member of the amino acid/polyamine/choline APC family, less than 10% of the numbers seen in *S. cerevisiae*, S. *pombe* or *Caenorhabditis elegans* (Fig. 5). The apparent lack of solute transporters in *P. falciparum* correlates with the lower percentage of multispanning membrane proteins compared with other eukaryotic organisms (Fig. 5). The predicted transport capabilities of *P. falciparum* resemble those of obligate intracellular prokaryotic parasites, which also possess a limited complement of transporters for organic solutes[102].

A complete catalogue of the identified transporters is presented in Fig. 4. In addition to the glucose/proton symporter[103] and the water/glycerol channel[104], one other probable sugar transporter and three carboxylate transporters were identified; one or more of the latter are probably responsible for the lactate and pyruvate/proton symport activity of *P. falciparum*[105]. Two nucleoside/nucleobase transporters are encoded on the *P. falciparum* genome, one of which has been localized to the parasite plasma membrane[106]. No obvious amino-acid transporters were detected, which emphasizes the importance of haemoglobin digestion within the food vacuole as an important source of amino acids for the erythrocytic stages of the parasite. How the insect stages of the parasite acquire amino acids and other important nutrients is unknown, but four metabolic uptake systems were identified whose substrate specificity could not be predicted with confidence. The parasite may also possess novel proteins that mediate these activities. Nine members of the mitochondrial carrier family are present in *P. falciparum*, including an ATP/ADP exchanger[107] and a di/tri-carboxylate exchanger, probably involved in transport of TCA cycle intermediates across the mitochondrial membrane. Probable phosphoenolpyruvate/phosphate and sugar phosphate/phosphate antiporters most similar to those of plant chloroplasts were identified, suggesting that these transporters are targeted to the apicoplast membrane. The former may enable uptake of phosphoenolpyruvate as a precursor of fatty-acid biosynthesis.

A more extensive set of transporters could be identified for the transport of inorganic ions and for export of drugs and hydrophobic compounds. Sodium/proton and calcium/proton exchangers were identified, as well as other metal cation transporters, including a substantial set of 16 P-type ATPases. An Nramp divalent cation transporter was identified which may be specific for manganese or iron. *Plasmodium falciparum* contains all subunits of V-type ATPases as well as two proton translocating pyrophosphatases[108], which could be used to generate a proton motive force, possibly across the parasite plasma membrane as well as across a vacuolar membrane. The proton pumping pyrophosphatases are not present in mammals, and could form attractive antimalarial targets. Only a single copy of the *P. falciparum* chloroquine-resistance gene *crt* is present, but multiple homologues of the multidrug resistance pump *mdr1* and other predicted multidrug transporters were identified (Fig. 3). Mutations in *crt* seem to have a central role in the development of chloroquine resistance[109].

*Plasmodium falciparum* infection of erythrocytes causes a variety of pleiotropic changes in host membrane transport. Patch clamp analysis has described a novel broad-specificity channel activated or inserted in the red blood cell membrane by *P. falciparum* infection that allows uptake of various nutrients[110]. If this channel is encoded by the parasite, it is not obvious from genome analysis, because no clear homologues of eukaryotic sodium, potassium or chloride ion channels could be identified. This suggests that *P. falciparum* may use one or more novel membrane channels for this activity.

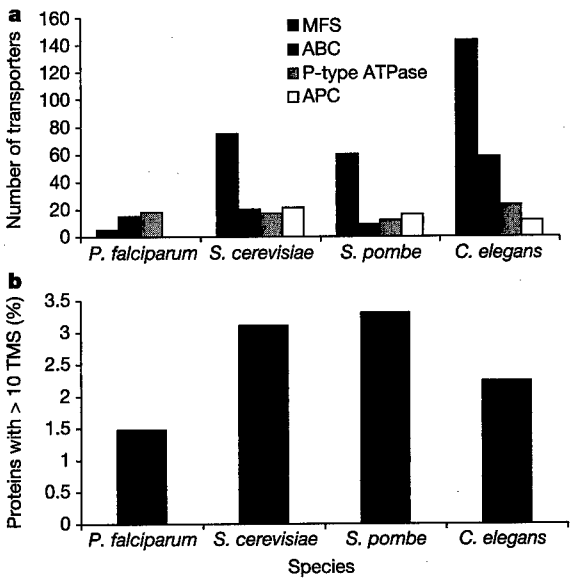



**Figure 5** Analysis of transporters in *P. falciparum*. **a**, Comparison of the numbers of transporters belonging to the major facilitator superfamily (MFS), ATP-binding cassette (ABC) family, P-type ATPase family and the amino acid/polyamine/choline (APC) family in *P. falciparum* and other eukaryotes. Analyses were performed as previously described[102]. **b**, Comparison of the numbers of proteins with ten or more predicted transmembrane segments[163] (TMS) in *P. falciparum* and other eukaryotes. Prediction of membrane spanning segments was performed using TMHMM.

## DNA replication, repair and recombination

DNA repair processes are involved in maintenance of genomic integrity in response to DNA damaging agents such as irradiation, chemicals and oxygen radicals, as well as errors in DNA metabolism such as misincorporation during DNA replication. The *P. falciparum* genome encodes at least some components of the major DNA repair processes that have been found in other eukaryotes[111,112]. The core of eukaryotic nucleotide excision repair is present (XPB/Rad25, XPG/Rad2, XPF/Rad1, XPD/Rad3, ERCC1) although some highly conserved proteins with more accessory roles

could not be found (for example, XPA/Rad4, XPC). The same is true for homologous recombinational repair with core proteins such as MRE11, DMC1, Rad50 and Rad51 present but accessory proteins such as NBS1 and XRS2 not yet found. These accessory proteins tend to be poorly conserved and have not been found outside of animals or yeast, respectively, and thus may be either absent or difficult to identify in *P. falciparum*. However, it is interesting that Archaea possess many of the core proteins but not the accessory proteins for these repair processes, suggesting that many of the accessory eukaryotic repair proteins evolved after *P. falciparum* diverged from other eukaryotes.

The presence of MutL and MutS homologues including possible orthologues of MSH2, MSH6, MLH1 and PMS1 suggests that *P. falciparum* can perform post-replication mismatch repair. Orthologues of MSH4 and MSH5, which are involved in meiotic crossing over in other eukaryotes, are apparently absent in *P. falciparum*. The repair of at least some damaged bases may be performed by the combined action of the four base excision repair glycosylase homologues and one of the apurinic/apyrimidinic (AP) endonucleases (homologues of Xth and Nfo are present). Experimental evidence suggests that this is done by the long-patch pathway[113].

The presence of a class II photolyase homologue is intriguing, because it is not clear whether *P. falciparum* is exposed to significant amounts of ultraviolet irradiation during its life cycle. It is possible that this protein functions as a blue-light receptor instead of a photolyase, as do members of this gene family in some organisms such as humans. Perhaps most interesting is the apparent absence of homologues of any of the genes encoding enzymes known to be involved in non-homologous end joining (NHEJ) in eukaryotes (for example, Ku70, Ku86, Ligase IV and XRCC1)[112]. NHEJ is involved in the repair of double strand breaks induced by irradiation and chemicals in other eukaryotes (such as yeast and humans), and is also involved in a few cellular processes that create double strand breaks (for example, VDJ recombination in the immune system in humans). The role of NHEJ in repairing radiation-induced double strand breaks varies between species[114]. For example, in humans, cells with defects in NHEJ are highly sensitive to γ-irradiation while yeast mutants are not. Double strand breaks in yeast are repaired primarily by homologous recombination. As NHEJ is involved in regulating telomere stability in other organisms, its apparent absence in *P. falciparum* may explain some of the unusual properties of the telomeres in this species[115].

## Secretory pathway

*Plasmodium falciparum* contains genes encoding proteins that are important in protein transport in other eukaryotic organisms, but the organelles associated with a classical secretory pathway and protein transport are difficult to discern at an ultra-structural level[116]. In order to identify additional proteins that may have a role in protein translocation and secretion, the *P. falciparum* protein database was searched with *S. cerevisiae* proteins with GO assignments for involvement in protein export. We identified potential homologues of important components of the signal recognition particle, the translocon, the signal peptidase complex and many components that allow vesicle assembly, docking and fusion, such as COPI and COPII, clathrin, adaptin, v- and t-SNARE and GTP binding proteins. The presence of Sec62 and Sec63 orthologues raises the possibility of post-translational translocation of proteins, as found in *S. cerevisiae*.

Although *P. falciparum* contains many of the components associated with a classical secretory system and vesicular transport of proteins, the parasite secretory pathway has unusual features. The parasite develops within a parasitophorous vacuole that is formed during the invasion of the host cell, and the parasite modifies the host erythrocyte by the export of parasite-encoded proteins[117]. The mechanism(s) by which these proteins, some of which lack signal peptide sequences, are transported through and targeted beyond the membrane of the parasitophorous vacuole remains unknown. But these mechanisms are of particular importance because many of the proteins that contribute to the development of severe disease are exported to the cytoplasm and plasma membrane of infected erythrocytes.

Attempts to resolve these observations resulted in the proposal of a secondary secretory pathway[118]. More recent studies suggest export of COPII vesicle coat proteins, Sar1 and Sec31, to the erythrocyte cytoplasm as a mechanism of inducing vesicle formation in the host cell, thereby targeting parasite proteins beyond the parasitophorous vacuole, a new model in cell biology[119,120]. A homologue of *N*-ethylmaleimide-sensitive factor (NSF), a component of vesicular transport, has also been located to the erythrocyte cytoplasm[121]. The 41-2 antigen of *P. falciparum*, which is also found in the erythrocyte cytoplasm and plasma membrane[122], is homologous with BET3, a subunit of the *S. cerevisiae* transport protein particle (TRAPP) that mediates endoplasmic reticulum to Golgi vesicle docking and fusion[123]. It is not clear how these proteins are targeted to the cytoplasm, as they lack an obvious signal peptide. Nevertheless, the expanded list of protein-transport-associated genes identified in the *P. falciparum* genome should facilitate the development of specific probes to further elucidate the intra- and extracellular compartments of its protein transport system.

## Immune evasion

In common with other organisms, highly variable gene families are clustered towards the telomeres. *Plasmodium falciparum* contains three such families termed *var*, *rif* and *stevor*, which code for proteins known as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), repetitive interspersed family (rifin) and sub-telomeric variable open reading frame (stevor), respectively[5,124–130]. The 3D7 genome contains 59 *var*, 149 *rif* and 28 *stevor* genes, but for each family there are also a number of pseudogenes and gene truncations present.

The *var* genes code for proteins which are exported to the surface of infected red blood cells where they mediate adherence to host endothelial receptors[131], resulting in the sequestration of infected cells in a variety of organs. These and other adherence properties[132–135] are important virulence factors that contribute to the development of severe disease. Rifins, products of the *rif* genes, are also expressed on the surface of infected red cells and undergo antigenic variation[131]. Proteins encoded by *stevor* genes show sequence similarity to rifins, but they are less polymorphic than the rifins[129]. The function of rifins and stevors is unknown. PfEMP1 proteins are targets of the host protective antibody response[136], but transcriptional switching between *var* genes permits antigenic variation and a means of immune evasion, facilitating chronic infection and transmission. Products of the *var* gene family are thus central to the pathogenesis of malaria and to the induction of protective immunity.

Figure 6 shows the genome-wide arrangement of these multigene families. In the 24 chromosomal ends that have a *var* gene as the first transcriptional unit, there are three basic types of gene arrangement. Eight have the general pattern *var-rif var* + /− $(rif/stevor)_n$, ten can be described as *var*-$(rif/stevor)_n$, three have a *var* gene alone and two have two or more adjacent *var* genes. This telomeric organization is consistent with exchange between chromosome ends, although the extent of this re-assortment may be limited by the varied gene combinations. The *var*, *rif* and *stevor* genes consist of two exons. The first *var* exon is between 3.5 and 9.0 kb in length, polymorphic and encodes an extracellular region of the protein. The second exon is between 1.0 and 1.5 kb, and encodes a conserved cytoplasmic tail that contains acidic amino-acid residues (ATS; 'acidic terminal sequence'). The first *rif* and *stevor* exons are about 50–75 bp in length, and encode a putative signal sequence while the second exon is about 1 kb in length, with the *rif* exon being on average slightly larger than that for *stevor*. The rifin sequences fall into two major

subgroups determined by the presence or absence of a consensus peptide sequence, KEL ($X_{15}$) IPTCVCR, approximately 100 amino acids from the N terminus. The *var* genes are made up of three recognizable domains known as 'Duffy binding like' (DBL); 'cysteine rich interdomain region' (CIDR) and 'constant2' (C2)[137–139]. Alignment of sequences existing before the *P. falciparum* genome project had placed each of these domains into a number of sub-classes; α to ε for DBL domains, and α to γ for CIDR domains. Despite these recognizable signatures, there is a low level of sequence similarity even between domains of the same sub-type. Alignment and tree construction of the DBL domains identified here showed that a small number did not fit well into existing categories, and have been termed DBL-X. Similar analysis of all 3D7 CIDR sequences showed that with this data they were best described as CIDRα or CIDR non-α, as distinct tree branches for the other domain types were not observed. In terms of domain type and order, 16 types of *var* gene sequences were identified in this study.

Type 1 *var* genes, consisting of DBLα, CIDRα, DBLδ, and CIDR non-α followed by the ATS, are the most common structures, with 38 genes in this category (Fig. 6b). A total of 58 *var* genes commence with a DBLα domain, and in 51 cases this is followed by CIDRα, and in 46 *var* genes the last domain of the first exon is CIDR non-α. Four *var* genes are atypical with the first exon consisting solely of DBL domains (type 3 and type 13). There is non-randomness in the ordering and pairing of DBL and CIDR sub-domains[140], suggesting that some—for example, DBLδ–CIDR non-α and DBLβ–C2 (Table 3)—should either be considered as functional–structural combinations, or that recombination in these areas is not favoured, thereby preserving the arrangement. Eighteen of the 24 telomeric proximal *var* genes are of type 1. With two exceptions, type 4 on chromosome 7 and type 9 on chromosome 11, all of the telomeric *var* genes are transcribed towards the centromere. The inverted position of the two *var* genes may hinder homologous recombination at these loci in telomeric clusters that are formed during asexual multiplication[115]. A further 12 *var* genes are located near to telomeres, with the remaining *var* genes forming internal clusters on chromosomes 4, 7, 8 and 12 and a single internal gene being located on chromosome 6.

Alignment of sequences 1.5 kb upstream of all of the *var* genes revealed three classes of sequences, upsA, upsB and upsC (of which there are 11, 35 and 13 members, respectively) that show preferential association with different *var* genes. Thus, upsB is associated with 22 out of 24 telomeric *var* genes, upsA is found with the two remaining telomeric *var* genes that are transcribed towards the telomere and with most telomere associated *var* genes (9 out of 12) which also point towards the telomere[141]. All 13 upsC sequences are associated with internal *var* clusters. Nearly all the telomeric *var* genes have an (A + T)-rich region approximately 2 kb upstream characterized by a number of poly(A) tracts as well as one or more copies of the consensus GGATCTAG. An analysis of the regions 1.0 kb downstream of *var* genes shows three sequence families, with members of one family being associated primarily with *var* genes

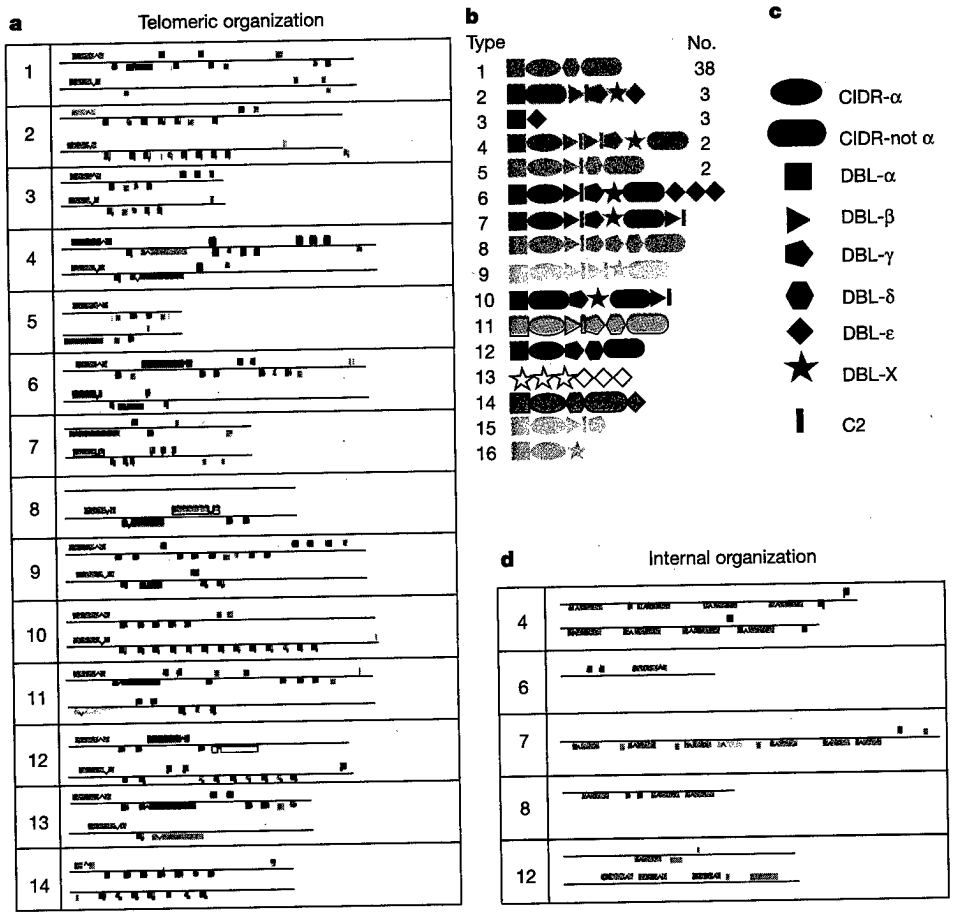



**Figure 6** Organization of multi-gene families in *P. falciparum*. **a**, Telomeric regions of all chromosomes showing the relative positions of members of the multi-gene families: *rif* (blue) *stevor* (yellow) and *var* (colour coded as indicated; see **b** and **c**). Grey boxes represent pseudogenes or gene fragments of any of these families. The left telomere is shown above the right. Scale: ~0.6 mm = 1 kb. **b**, **c**, *var* gene domain structure. *var* genes contain three domain types: DBL, of which there are six sequence classes; CIDR, of which there are two sequence classes; and conserved 2 (C2) domains (see text). The relative order of the domains in each gene is indicated (**c**). *var* genes with the same domain types in the same order have been colour coded as an identical class and given an arbitrary number for their type (**b**) and the total number of members of each class in the genome of *P. falciparum* clone 3D7. **d**, Internal multi-gene family clusters. Key as in **a**.

next to the telomeric repeats. The intron sequences within the *var* genes have been associated with locus specific silencing[142]. They vary in length from 170 to ~1,200 bp and are ~89% A/T. On the coding strand, at the 5′ end the non-A/T bases are mainly G residues with 70% of sequences having the consensus TGTTTGGATATATA. The central regions are highly A-rich, and contain a number of semi-conserved motifs. The 3′ region is comparably rich in C, with one or more copies in most genes of the sequence $(TA)_n$ CCCATAAC-TACA. The 3′ end has an extended and atypical splice consensus of ACANATATAGTTA$(T)_n$ TAG. Sequences upstream of *rif* and *stevor* genes also have distinguishable upstream sequences, but a proportion of *rif* genes have the *stevor* type of 5′ sequence. Because the majority of telomeric *var* genes share a similar structure and 5′ and 3′ sequences, they may form a unique group in terms of regulation of gene expression.

The most conserved *var* gene previously identified, which mediates adherence to chondroitin sulphate A in the placenta[143], is incomplete in 3D7 because of deletion of part of exon 1 and all of exon 2. This gene is located on the right telomere of chromosome 5 (Fig 6). The majority of *var* genes sequenced previously had been identified as they mediated adhesion to particular receptors, and most of them had more than four domains in exon 1. The fact that type 1 *var* genes containing only 4 domains predominate in the 3D7 genome suggests that previous analyses had been based on a highly biased sample. The significance of this in terms of the function of type 1 *var* genes remains to be determined.

Immune-evasion mechanisms such as clonal antigenic variation of parasite-derived red cell surface proteins (PfEMP1s, rifins) and modulation of dendritic cell function have been documented in *P. falciparum*[131,132]. A putative homologue of human cytokine macrophage migration inhibitory factor (MIF) was identified in *P. falciparum*. In vertebrates, MIFs have been shown to function as immuno-modulators and as growth factors[144], and in the nematode *Brugia malayi*, recombinant MIF modulated macrophage migration and promoted parasite survival[145]. An MIF-type protein in *P. falciparum* may contribute to the parasite's ability to modulate the immune response by molecular mimicry or participate in other host–parasite interactions.

## Implications for vaccine development

An effective malaria vaccine must induce protective immune responses equivalent to, or better than, those provided by naturally acquired immunity or immunization with attenuated sporozoites[146]. To date, about 30 *P. falciparum* antigens that were identified via conventional techniques are being evaluated for use in vaccines, and several have been tested in clinical trials. Partial protection with one vaccine has recently been attained in a field setting[147]. The present genome sequence will stimulate vaccine development by the identification of hundreds of potential antigens that could be scanned for desired properties such as surface expression or limited antigenic diversity. This could be combined with data on stage-specific expression obtained by microarray and proteomics[14,15] analyses to identify potential antigens that are expressed in one or more stages of the life cycle. However, high-throughput immunological assays to identify novel candidate vaccine antigens that are the targets of protective humoral and cellular immune responses in humans need to be developed if the genome sequence is to have an impact on vaccine development. In addition, new methods for maximizing the magnitude, quality and longevity of protective immune responses will be required in order to produce effective malaria vaccines.

Table 3 **Domains of PfEMP1 proteins in *P. falciparum***

| Domain type | Number of domains |
|---|---|
| DBLα | 58 |
| DBLβ-C2 | 18 |
| DBLγ | 13 |
| DBLδ | 44 |
| DBLε | 13 |
| DBL-X | 13 |
| CIDRα | 51 |
| CIDR non-α | 54 |
| **Preferred pairings** | **Frequency** |
| DBLα–CIDRα | 51/58 |
| DBLβ–C2 | 18/18 |
| DBLδ–CIDR non-α | 44/44 |
| CIDRα–DBLδ | 39/51 |
| CIDRα–DBLβ | 10/51 |
| DBLβ–C2–DBLγ | 10/18 |
| DBLγ–DBL-X | 8/13 |

Top, the total number of each DBL or CIDR domain type in intact *var* genes within the *P. falciparum* 3D7 genome. Bottom, the frequencies of the most common individual domain pairings found within intact *var* genes. The denominator refers to the total number of the first-named domains in intact *var* genes, and the numerator refers to the number of second-named domains found adjacent. See text for discussion of domain types.

## Concluding remarks

The *P. falciparum*, *Anopheles gambiae* and *Homo sapiens* genome sequences have been completed in the past two years, and represent new starting points in the centuries-long search for solutions to the malaria problem. For the first time, a wealth of information is available for all three organisms that comprise the life cycle of the malaria parasite, providing abundant opportunities for the study of each species and their complex interactions that result in disease. The rapid pace of improvements in sequencing technology and the declining costs of sequencing have made it possible to begin genome sequencing efforts for *Plasmodium vivax*, the second major human malaria parasite, several malaria parasites of animals, and for many related parasites such as *Theileria* and *Toxoplasma*. These will be extremely useful for comparative purposes. Last, this technology will enable sampling of parasite, vector and host genomes in the field, providing information to support the development, deployment and monitoring of malaria control methods.

In the short term, however, the genome sequences alone provide little relief to those suffering from malaria. The work reported here and elsewhere needs to be accompanied by larger efforts to develop new methods of control, including new drugs and vaccines, improved diagnostics and effective vector control techniques. Much remains to be done. Clearly, research and investments to develop and implement new control measures are needed desperately if the social and economic impacts of malaria are to be relieved. The increased attention given to malaria (and to other infectious diseases affecting tropical countries) at the highest levels of government, and the initiation of programmes such as the Global Fund to Fight AIDS, Tuberculosis and Malaria[148], the Multilateral Initiative on Malaria in Africa[149], the Medicines for Malaria Venture[150], and the Roll Back Malaria campaign[151], provide some hope of progress in this area. It is our hope and expectation that researchers around the globe will use the information and biological insights provided by complete genome sequences to accelerate the search for solutions to diseases affecting the most vulnerable of the world's population. □

## Methods

### Sequencing, gap closure and annotation

The techniques used at each of the three participating centres for sequencing, closure and annotation are described in the accompanying Letters[7–9]. To ensure that each centres' annotation procedures produced roughly equivalent results, the Wellcome Trust Sanger Institute ('Sanger') and the Institute for Genomic Research ('TIGR') annotated the same 100-kb segment of chromosome 14. The number of genes predicted in this sequence by the two centres was 22 and 23; the discrepancy being due to the merging of two single genes by one centre. Of the 74 exons predicted by the two centres, 50 (68%) were identical, 9 (2%) overlapped, 6 (8%) overlapped and shared one boundary, and the remainder were predicted by one centre but not the other. Thus 88% of the exons predicted by the two centres in the 100-kb fragment were identical or overlapped.

Finished sequence data and annotation were transferred in XML (extensible markup language) format from Sanger and the Stanford Genome Technology Center to TIGR, and

made available to co-authors over the internet. Genes on finished chromosomes were assigned systematic names according the scheme described previously[5]. Genes on unfinished chromosomes were given temporary identifiers.

## Analysis of subtelomeric regions

Subtelomeric regions were analysed by the alignment of all of the chromosomes to each other using MUMmer2[152] with a minimum exact match length ranging from 30 to 50 bp. Tandem repeats were identified by extracting a 90-kb region from the ends of all chromosomes and using Tandem Repeat Finder[153] with the following parameter settings: match = 2, mismatch = 7, indel = 7, pm = 75, pi = 10, minscore = 100, maxperiod = 500. Detailed pairwise alignments of internal telomeric blocks were computed with the ssearch program from the Fasta3 package[154].

## Evolutionary analyses

*Plasmodium falciparum* proteins were searched against a database of proteins from all complete genomes as well as from a set of organelle, plasmid and viral genomes. Putative recently duplicated genes were identified as those encoding proteins with better BLASTP matches (based on E value with a $10^{-15}$ cutoff) to other proteins in *P. falciparum* than to proteins in any other species. Proteins of possible organellar descent were identified as those for which one of the top six prokaryotic matches (based on E value) was to either a protein encoded by an organelle genome or by a species related to the organelle ancestors (members of the *Rickettsia* subgroup of the α-Proteobacteria or cyanobacteria). Because BLAST matches are not an ideal method of inferring evolutionary history, phylogenetic analysis was conducted for all these proteins. For phylogenetic analysis, all homologues of each protein were identified by BLASTP searches of complete genomes and of a non-redundant protein database. Sequences were aligned using CLUSTALW, and phylogenetic trees were inferred using the neighbour-joining algorithms of CLUSTALW and PHYLIP. For comparative analysis of eukaryotes, the proteomes of all eukaryotes for which complete genomes are available (except the highly reduced *E. cuniculi*) were searched against each other. The proportion of proteins in each eukaryotic species that had a BLASTP match in each of the other eukaryotic species was determined, and used to infer a 'whole-genome tree' using the neighbour-joining algorithm. Possible eukaryotic conserved and specific proteins were identified as those with matches to all the complete eukaryotic genomes ($10^{-30}$ E-value cutoff) but without matches to any complete prokaryotic genome ($10^{-15}$ cutoff).

Received 31 July; accepted 2 September 2002; doi:10.1038/nature01097.


1. Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64,** 1–11 (2001).
2. Greenwood, B. & Mutabingwa, T. Malaria in 2002. *Nature* **415,** 670–672 (2002).
3. Gallup, J. L. & Sachs, J. D. The economic burden of malaria. *Am. J. Trop. Med. Hyg.* **64,** 85–96 (2001).
4. Hoffman, S. L. *et al.* Funding for malaria genome sequencing. *Nature* **387,** 647 (1997).
5. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
6. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
7. Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419,** 527–531 (2002).
8. Gardner, M. J. *et al.* Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419,** 531–534 (2002).
9. Hyman, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419,** 534–537 (2002).
10. Foster, J. & Thompson, J. The *Plasmodium falciparum* genome project: a resource for researchers. *Parasitol. Today* **11,** 1–4 (1995).
11. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
12. Su, X. Z. & Wellems, T. E. Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats. *Genomics* **33,** 430–444 (1996).
13. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
14. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).
15. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542 (2002).
16. Watanabe, J., Sasaki, M., Suzuki, Y. & Sugano, S. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*. *Nucleic Acids Res.* **29,** 70–71 (2001).
17. Gamain, B. *et al.* Increase in glutathione peroxidase activity in malaria parasite after selenium supplementation. *Free Radic. Biol. Med.* **21,** 559–565 (1996).
18. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450–453 (2001).
19. Moriyama, E. N. & Powell, J. R. Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* **45,** 514–523 (1997).
20. Duret, L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16,** 287–289 (2000).
21. Vaidya, A. B., Akella, R. & Suplick, K. Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malaria parasite. *Mol. Biochem. Parasitol.* **35,** 97–107 (1989).
22. Vaidya, A. B., Lashgari, M. S., Pologe, L. G. & Morrisey, J. Structural features of *Plasmodium* cytochrome *b* that may underlie susceptibility to 8-aminoquinolines and hydroxynaphthoquinones. *Mol. Biochem. Parasitol.* **58,** 33–42 (1993).
23. Tan, T. H., Pach, R., Crausaz, A., Ivens, A. & Schneider, A. tRNAs in *Trypanosoma brucei*: genomic organization, expression, and mitochondrial import. *Mol. Cell. Biol.* **22,** 3707–3717 (2002).
24. Tarassov, I. A. & Martin, R. P. Mechanisms of tRNA import into yeast mitochondria: an overview. *Biochimie* **78,** 502–510 (1996).
25. Wilson, R. J. M. *et al.* Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. *J. Mol. Biol.* **261,** 155–172 (1996).
26. Li, J., Wirtz, R. A., McConkey, G. A., Sattabongkot, J. & McCutchan, T. F. Transition of *Plasmodium vivax* ribosome types corresponds to sporozoite differentiation in the mosquito. *Mol. Biochem. Parasitol.* **65,** 283–289 (1994).
27. Waters, A. P. The ribosomal RNA genes of *Plasmodium*. *Adv. Parasitol.* **34,** 33–79 (1994).
28. Babiker, H. A., Creasey, A. M., Bayoumi, R. A., Walliker, D. & Arnot, D. E. Genetic diversity of *Plasmodium falciparum* in a village in eastern Sudan. 2. Drug resistance, molecular karyotypes and the mdr1 genotype of recent isolates. *Trans. R. Soc. Trop. Med. Hyg.* **85,** 578–583 (1991).
29. Hinterberg, K., Mattei, D., Wellems, T. E. & Scherf, A. Interchromosomal exchange of a large subtelomeric segment in a *Plasmodium falciparum* cross. *EMBO J.* **13,** 4174–4180 (1994).
30. Hernandez, R. R., Hinterberg, K. & Scherf, A. Compartmentalization of genes coding for immunodominant antigens to fragile chromosome ends leads to dispersed subtelomeric gene families and rapid gene evolution in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **78,** 137–148 (1996).
31. Scherf, A. *et al.* Gene inactivation of Pf11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametocytogenesis. *EMBO J.* **11,** 2293–2301 (1992).
32. Day, K. P. *et al.* Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. *Proc. Natl Acad. Sci. USA* **90,** 8292–8296 (1993).
33. Pologe, L. G. & Ravetch, J. V. A chromosomal rearrangement in a *P. falciparum* histidine-rich protein gene is associated with the knobless phenotype. *Nature* **322,** 474–477 (1986).
34. Louis, E. J., Naumova, E. S., Lee, A., Naumov, G. & Haber, J. E. The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. *Genetics* **136,** 789–802 (1994).
35. van Deutekom, J. C. *et al.* Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counselling and etiology of FSHD1. *Hum. Mol. Genet.* **5,** 1997–2003 (1996).
36. Rudenko, G., McCulloch, R., Dirks-Mulder, A. & Borst, P. Telomere exchange can be an important mechanism of variant surface glycoprotein gene switching in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **80,** 65–75 (1996).
37. Figueiredo, L. M., Freitas-Junior, L. H., Bottius, E., Olivo-Marin, J. C. & Scherf, A. A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J.* **21,** 815–824 (2002).
38. Scherf, A., Figueiredo, L. M. & Freitas-Junior, L. H. *Plasmodium* telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.* **4,** 409–414 (2001).
39. Vernick, K. D. & McCutchan, T. F. Sequence and structure of a *Plasmodium falciparum* telomere. *Mol. Biochem. Parasitol.* **28,** 85–94 (1988).
40. Oquendo, P. *et al.* Characterisation of a repetitive DNA sequence from the malaria parasite, *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **18,** 89–101 (1986).
41. De Bruin, D., Lanzer, M. & Ravetch, J. V. The polymorphic subtelomeric regions of *Plasmodium falciparum* chromosomes contain arrays of repetitive sequence elements. *Proc. Natl Acad. Sci. USA* **91,** 619–623 (1994).
42. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
43. McFadden, G. I., Reith, M., Munhollan, J. & Lang-Unnasch, N. Plastid in human parasites. *Nature* **381,** 482–483 (1996).
44. Kohler, S. *et al.* A plastid of probable green algal origin in apicomplexan parasites. *Science* **275,** 1485–1489 (1997).
45. Fichera, M. E. & Roos, D. S. A plastid organelle as a drug target in apicomplexan parasites. *Nature* **390,** 407–409 (1997).
46. He, C. Y., Striepen, B., Pletcher, C. H., Murray, J. M. & Roos, D. S. Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of *Toxoplasma gondii*. *J. Biol. Chem.* **276,** 28436–28442 (2001).
47. Waller, R. F. *et al.* Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **95,** 12352–12357 (1998).
48. Surolia, N. & Surolia, A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. *Nature Med.* **7,** 167–173 (2001).
49. Jomaa, H. *et al.* Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285,** 1573–1576 (1999).
50. Sato, S. & Wilson, R. J. The genome of *Plasmodium falciparum* encodes an active delta-aminolevulinic acid dehydratase. *Curr. Genet.* **40,** 391–398 (2002).
51. Van Dooren, G. G., Su, V., D'Ombrain, M. C. & McFadden, G. I. Processing of an apicoplast leader sequence in *Plasmodium falciparum* and the identification of a putative leader cleavage enzyme. *J. Biol. Chem.* **277,** 23612–23619 (2002).
52. Wilson, R. J. Progress with parasite plastids. *J. Mol. Biol.* **319,** 257–274 (2002).
53. Stoebe, B. & Kowallik, K. V. Gene-cluster analysis in chloroplast genomics. *Trends Genet.* **15,** 344–347 (1999).
54. Fast, N. M., Kissinger, J. C., Roos, D. S. & Keeling, P. J. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol. Biol. Evol* **18,** 418–426 (2001).
55. Roos, D. S. *et al.* Origin, targeting, and function of the apicomplexan plastid. *Curr. Opin. Microbiol.* **2,** 426–432 (1999).
56. Palmer, J. D. & Delwiche, C. F. Second-hand chloroplasts and the case of the disappearing nucleus. *Proc. Natl Acad. Sci. USA* **93,** 7432–7435 (1996).
57. Waller, R. F., Reed, M. B., Cowman, A. F. & McFadden, G. I. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *EMBO J.* **19,** 1794–1802 (2000).
58. DeRocher, A., Hagen, C. B., Froehlich, J. E., Feagin, J. E. & Parsons, M. Analysis of targeting sequences demonstrates that trafficking to the *Toxoplasma gondii* plastid branches off the secretory system. *J. Cell Sci.* **113** (Part 22), 3969–3977 (2000).
59. van Dooren, G. G., Schwartzbach, S. D., Osafune, T. & McFadden, G. I. Translocation of proteins across the multiple membranes of complex plastids. *Biochim. Biophys. Acta* **1541,** 34–53 (2001).

60. Yung, S., Unnasch, T. R. & Lang-Unnasch, N. Analysis of apicoplast targeting and transit peptide processing in *Toxoplasma gondii* by deletional and insertional mutagenesis. *Mol. Biochem. Parasitol.* **118,** 11–21 (2001).
61. Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I. & Schneider, G. Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280,** 19–26 (2001).
62. Vollmer, M., Thomsen, N., Wiek, S. & Seeber, F. Apicomplexan parasites possess distinct nuclear-encoded, but apicoplast-localized, plant-type ferredoxin-NADP$^+$ reductase and ferredoxin. *J. Biol. Chem.* **276,** 5483–5490 (2001).
63. Ralph, S. A., D'Ombrain, M. C. & McFadden, G. I. The apicoplast as an antimalarial drug target. *Drug Resist. Updat.* **4,** 145–151 (2001).
64. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).
65. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
66. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8,** 163–167 (1998).
67. Adams, K. L., Daley, D. O., Whelan, J. & Palmer, J. D. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. *Plant Cell* **14,** 931–943 (2002).
68. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30,** 42–46 (2002).
69. Sherman, I. W. in *Malaria Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 135–143 (ASM, Washington DC, 1998).
70. Buckwitz, D., Jacobasch, G., Gerth, C., Holzhutter, H. G. & Thamm, R. A kinetic model of phosphofructokinase from *Plasmodium berghei*. Influence of ATP and fructose-6-phosphate. *Mol. Biochem. Parasitol.* **27,** 225–232 (1988).
71. Buckwitz, D., Jacobasch, G. & Gerth, C. Phosphofructokinase from *Plasmodium berghei*. Influence of $Mg^{2+}$, ATP and $Mg^{2+}$-complexed ATP. *Biochem. J.* **267,** 353–357 (1990).
72. Clarke, J. L., Scopes, D. A., Sodeinde, O. & Mason, P. J. Glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase. A novel bifunctional enzyme in malaria parasites. *Eur. J. Biochem.* **268,** 2013–2019 (2001).
73. Miclet, E. *et al.* NMR spectroscopic analysis of the first two steps of the pentose-phosphate pathway elucidates the role of 6-phosphogluconolactonase. *J. Biol. Chem.* **276,** 34840–34846 (2001).
74. Loyevsky, M. *et al.* An IRP-like protein from *Plasmodium falciparum* binds to a mammalian iron-responsive element. *Blood* **98,** 2555–2562 (2001).
75. Lang-Unnasch, N. Purification and properties of *Plasmodium falciparum* malate dehydrogenase. *Mol. Biochem. Parasitol.* **50,** 17–25 (1992).
76. Blum, J. J. & Ginsburg, H. Absence of α-ketoglutarate dehydrogenase activity and presence of $CO_2$-fixing activity in *Plasmodium falciparum* grown *in vitro* in human erythrocytes. *J. Protozool.* **31,** 167–169 (1984).
77. Fry, M. & Beesley, J. E. Mitochondria of mammalian *Plasmodium* spp. *Parasitology* **102,** 17–26 (1991).
78. Vaidya, A. B. in *Malaria: Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 355–368 (ASM, Washington DC, 1998).
79. Papa, S., Zanotti, F. & Gaballo, A. The structural and functional connection between the catalytic and proton translocating sectors of the mitochondrial $F_1F_0$-ATP synthase. *J. Bioenerg. Biomembr.* **32,** 401–411 (2000).
80. Sherman, I. W. in *Malaria: Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 177–184 (ASM, Washington DC, 1998).
81. de Macedo, C. S., Uhrig, M. L., Kimura, E. A. & Katzin, A. M. Characterization of the isoprenoid chain of coenzyme Q in *Plasmodium falciparum*. *FEMS Microbiol. Lett.* **207,** 13–20 (2002).
82. Trumpower, B. L. & Gennis, R. B. Energy transduction by cytochrome complexes in mitochondrial and bacterial respiration: the enzymology of coupling electron transfer reactions to transmembrane proton translocation. *Annu. Rev. Biochem.* **63,** 675–716 (1994).
83. Vaidya, A. B., McIntosh, M. T. & Srivastava, I. K. *Membrane Structure in Disease and Drug Therapy* (ed. Zimmer, G.) (Marcel Dekker, New York, 2000).
84. Perez-Martinez, X. *et al.* Subunit II of cytochrome c oxidase in Chlamydomonad algae is a heterodimer encoded by two independent nuclear genes. *J. Biol. Chem.* **276,** 11302–11309 (2001).
85. Murphy, A. D. & Lang-Unnasch, N. Alternative oxidase inhibitors potentiate the activity of atovaquone against *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **43,** 651–654 (1999).
86. Dieckmann, A. & Jung, A. Mechanisms of sulfadoxine resistance in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **19,** 143–147 (1986).
87. McConkey, G. A. Targeting the shikimate pathway in the malaria parasite *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **43,** 175–177 (1999).
88. Roberts, F. *et al.* Evidence for the shikimate pathway in apicomplexan parasites. *Nature* **393,** 801–805 (1998).
89. Roberts, C. W. *et al.* The shikimate pathway and its branches in apicomplexan parasites. *J. Infect. Dis.* **185** (Suppl. 1), S25–S36 (2002).
90. Keeling, P. J. *et al.* Shikimate pathway in apicomplexan parasites. *Nature* **397,** 219–220 (1999).
91. Fitzpatrick, T. *et al.* Subcellular localization and characterization of chorismate synthase in the apicomplexan *Plasmodium falciparum*. *Mol. Microbiol.* **40,** 65–75 (2001).
92. Duncan, K., Edwards, R. M. & Coggins, J. R. The pentafunctional arom enzyme of *Saccharomyces cerevisiae* is a mosaic of monofunctional domains. *Biochem. J.* **246,** 375–386 (1987).
93. Rubin, H. *et al.* Cloning, sequence determination, and regulation of the ribonucleotide reductase subunits from *Plasmodium falciparum*: a target for antimalarial therapy. *Proc. Natl Acad. Sci. USA* **90,** 9280–9284 (1993).
94. Chakrabarti, D., Schuster, S. M. & Chakrabarti, R. Cloning and characterization of subunit genes of ribonucleotide reductase, a cell-cycle-regulated enzyme, from *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **90,** 12020–12024 (1993).
95. Krnajski, Z., Gilberger, T. W., Walter, R. D. & Muller, S. The malaria parasite *Plasmodium falciparum* possesses a functional thioredoxin system. *Mol. Biochem. Parasitol.* **112,** 219–228 (2001).
96. Bonday, Z. Q., Dhanasekaran, S., Rangarajan, P. N. & Padmanaban, G. Import of host δ-aminolevulinate dehydratase into the malarial parasite: identification of a new drug target. *Nature Med.* **6,** 898–903 (2000).
97. Bonday, Z. Q., Taketani, S., Gupta, P. D. & Padmanaban, G. Heme biosynthesis by the malarial parasite. Import of δ-aminolevulinate dehydrase from the host red cell. *J. Biol. Chem.* **272,** 21839–21846 (1997).
98. Wilson, C. M., Smith, A. B. & Baylon, R. V. Characterization of the δ-aminolevulinate synthase gene homologue in *P. falciparum*. *Mol. Biochem. Parasitol.* **75,** 271–276 (1996).
99. Sato, S., Tews, I. & Wilson, R. J. Impact of a plastid-bearing endocytobiont on apicomplexan genomes. *Int. J. Parasitol.* **30,** 427–439 (2000).
100. Rohdich, F. *et al.* Biosynthesis of terpenoids. 2C-Methyl-D-erythritol 2,4-cyclodiphosphate synthase (IspF) from *Plasmodium falciparum*. *Eur. J. Biochem.* **268,** 3190–3197 (2001).
101. Kemp, L. E., Bond, C. S. & Hunter, W. N. Structure of 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase: an essential enzyme for isoprenoid biosynthesis and target for antimicrobial drug development. *Proc. Natl Acad. Sci. USA* **99,** 6591–6596 (2002).
102. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301,** 75–100 (2000).
103. Woodrow, C. J., Burchmore, R. J. & Krishna, S. Hexose permeation pathways in *Plasmodium falciparum*-infected erythrocytes. *Proc. Natl Acad. Sci. USA* **97,** 9931–9936 (2000).
104. Hansen, M., Kun, J. F., Schultz, J. E. & Beitz, E. A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites. *J. Biol. Chem.* **277,** 4874–4882 (2002).
105. Elliott, J. L., Saliba, K. J. & Kirk, K. Transport of lactate and pyruvate in the intraerythrocytic malaria parasite, *Plasmodium falciparum*. *Biochem. J.* **355,** 733–739 (2001).
106. Rager, N., Mamoun, C. B., Carter, N. S., Goldberg, D. E. & Ullman, B. Localization of the *Plasmodium falciparum* PfNT1 nucleoside transporter to the parasite plasma membrane. *J. Biol. Chem.* **276,** 41095–41099 (2001).
107. Dyer, M., Wong, I. H., Jackson, M., Huynh, P. & Mikkelsen, R. Isolation and sequence analysis of a cDNA encoding an adenine nucleotide translocator from *Plasmodium falciparum*. *Biochim. Biophys. Acta* **1186,** 133–136 (1994).
108. McIntosh, M. T., Drozdowicz, Y. M., Laroiya, K., Rea, P. A. & Vaidya, A. B. Two classes of plant-like vacuolar-type $H^+$-pyrophosphatases in malaria parasites. *Mol. Biochem. Parasitol.* **114,** 183–195 (2001).
109. Fidock, A. D. *et al.* Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* **6,** 861–871 (2000).
110. Desai, S. A., Bezrukov, S. M. & Zimmerberg, J. A voltage-dependent channel involved in nutrient uptake by red blood cells infected with the malaria parasite. *Nature* **406,** 1001–1005 (2000).
111. Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435,** 171–213 (1999).
112. Wood, R. D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291,** 1284–1289 (2001).
113. Haltiwanger, B. M. *et al.* DNA base excision repair in human malaria parasites is predominantly by a long-patch pathway. *Biochemistry* **39,** 763–772 (2000).
114. Critchlow, S. E. & Jackson, S. P. DNA end-joining: from yeast to man. *Trends Biochem. Sci.* **23,** 394–398 (1998).
115. Freitas-Junior, L. H. *et al.* Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407,** 1018–1022 (2000).
116. Bannister, L. H., Hopkins, J. M., Fowler, R. E., Krishna, S. & Mitchell, G. H. A brief illustrated guide to the ultrastructure of *Plasmodium falciparum* asexual blood stages. *Parasitol. Today* **16,** 427–433 (2000).
117. van Dooren, G. G., Waller, R. F., Joiner, K. A., Roos, D. S. & McFadden, G. I. Traffic jams: protein transport in *Plasmodium falciparum*. *Parasitol. Today* **16,** 421–427 (2000).
118. Wiser, M. F., Lanners, H. N., Bafford, R. A. & Favaloro, J. M. A novel alternate secretory pathway for the export of *Plasmodium* proteins into the host erythrocyte. *Proc. Natl Acad. Sci. USA* **94,** 9108–9113 (1997).
119. Albano, F. R. *et al.* A homologue of Sar1p localises to a novel trafficking pathway in malaria-infected erythrocytes. *Eur. J. Cell Biol.* **78,** 453–462 (1999).
120. Adisa, A., Albano, F. R., Reeder, J., Foley, M. & Tilley, L. Evidence for a role for a *Plasmodium falciparum* homologue of Sec31p in the export of proteins to the surface of malaria parasite-infected erythrocytes. *J. Cell Sci.* **114,** 3377–3386 (2001).
121. Hayashi, M. *et al.* A homologue of N-ethylmaleimide-sensitive factor in the malaria parasite *Plasmodium falciparum* is exported and localized in vesicular structures in the cytoplasm of infected erythrocytes in the brefeldin A-sensitive pathway. *J. Biol. Chem.* **276,** 15249–15255 (2001).
122. Knapp, B., Hundt, E. & Kupper, H. A. A new blood stage antigen of *Plasmodium falciparum* transported to the erythrocyte surface. *Mol. Biochem. Parasitol.* **37,** 47–56 (1989).
123. Sacher, M. *et al.* TRAPP, a highly conserved novel complex on the cis-Golgi that mediates vesicle docking and fusion. *EMBO J.* **17,** 2494–2503 (1998).
124. Leech, J. H., Barnwell, J. W., Miller, L. H. & Howard, R. J. Identification of a strain-specific malarial antigen exposed on the surface of *Plasmodium falciparum*-infected erythrocytes. *J. Exp. Med.* **159,** 1567–1575 (1984).
125. Weber, J. L. Interspersed repetitive DNA from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **29,** 117–124 (1988).
126. Su, Z. *et al.* The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82,** 89–100 (1995).
127. Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82,** 77–87 (1995).
128. Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82,** 101–110 (1995).
129. Cheng, Q. *et al.* stevor and rif are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Mol. Biochem. Parasitol.* **97,** 161–176 (1998).
130. Kyes, S. A., Rowe, J. A., Kriek, N. & Newbold, C. I. Rifins: A second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **96,** 9333–9338 (1999).
131. Kyes, S., Horrocks, P. & Newbold, C. Antigenic variation at the infected red cell surface in malaria. *Annu. Rev. Microbiol.* **55,** 673–707 (2001).
132. Urban, B. C. *et al. Plasmodium falciparum*-infected erythrocytes modulate the maturation of dendritic cells. *Nature* **400,** 73–77 (1999).
133. Pain, A. *et al.* Platelet-mediated clumping of *Plasmodium falciparum*-infected erythrocytes is a common adhesive phenotype and is associated with severe malaria. *Proc. Natl Acad. Sci. USA* **98,** 1805–1810 (2001).

134. Fried, M. & Duffy, P. E. Adherence of *Plasmodium falciparum* to chondroitin sulfate A in the human placenta. *Science* **272,** 1502–1504 (1996).
135. Udomsangpetch, R. *et al. Plasmodium falciparum*-infected erythrocytes form spontaneous erythrocyte rosettes. *J. Exp. Med.* **169,** 1835–1840 (1989).
136. Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4,** 358–360 (1998).
137. Peterson, D. S., Miller, L. H. & Wellems, T. E. Isolation of multiple sequences from the *Plasmodium falciparum* genome that encode conserved domains homologous to those in erythrocyte binding proteins. *Proc. Natl Acad. Sci. USA* **92,** 7100–7104 (1995).
138. Baruch, D. I. *et al.* Identification of a region of PfEMP1 that mediates adherence of *Plasmodium falciparum* infected erythrocytes to CD36: conserved function with variant sequence. *Blood* **90,** 3766–3775 (1997).
139. Smith, J. D., Gamain, B., Baruch, D. I. & Kyes, S. Decoding the language of *var* genes and *Plasmodium falciparum* sequestration. *Trends Parasitol.* **17,** 538–545 (2001).
140. Smith, J. D. *et al.* Identification of a *Plasmodium falciparum* intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria. *Proc. Natl Acad. Sci. USA* **97,** 1766–1771 (2000).
141. Voss, T. S. *et al.* Genomic distribution and functional characterisation of two distinct and conserved *Plasmodium falciparum var* gene 5′ flanking sequences. *Mol. Biochem. Parasitol.* **107,** 103–115 (2000).
142. Deitsch, K. W., Calderwood, M. S. & Wellems, T. E. Malaria. Cooperative silencing elements in *var* genes. *Nature* **412,** 875–876 (2001).
143. Rowe, J. A., Kyes, S. A., Rogerson, S. J., Babiker, H. A. & Raza, A. Identification of a conserved *Plasmodium falciparum var* gene implicated in malaria in pregnancy. *J. Infect. Dis.* **185,** 1207–1211 (2002).
144. Lue, H., Kleemann, R., Calandra, T., Roger, T. & Bernhagen, J. Macrophage migration inhibitory factor (MIF): mechanisms of action and role in disease. *Microbes Infect.* **4,** 449–460 (2002).
145. Pastrana, D. V. *et al.* Filarial nematode parasites secrete a homologue of the human cytokine macrophage migration inhibitory factor. *Infect. Immun.* **66,** 5955–5963 (1998).
146. Richie, T. L. & Saul, A. Progress and challenges for malaria vaccines. *Nature* **415,** 694–701 (2002).
147. Bojang, K. A. *et al.* Efficacy of RTS,S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet* **358,** 1927–1934 (2001).
148. Kapp, C. Global fund on AIDS, tuberculosis, and malaria holds first board meeting. *Lancet* **359,** 414 (2002).
149. Nchinda, T. C. Malaria: a reemerging disease in Africa. *Emerg. Infect. Dis.* **4,** 398–403 (1998).
150. Ridley, R. G. Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature* **415,** 686–693 (2002).
151. Nabarro, D. N. & Tayler, E. M. The "roll back malaria" campaign. *Science* **280,** 2067–2068 (1998).
152. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30,** 2478–2483 (2002).
153. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
154. Pearson, W. R. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132,** 185–219 (2000).
155. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418,** 79–85 (2002).
156. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M. A. & Barrell, B. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp. Funct. Genom.* **2,** 143–154 (2001).
157. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
158. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).
159. Scharfe, C. *et al.* MITOP, the mitochondrial proteome database: 2000 update. *Nucleic Acids Res.* **28,** 155–158 (2000).
160. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786 (1996).
161. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
162. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).
163. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
164. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6 (1997).
165. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium Yoelii yoelii*. *Nature* **419,** 512–519 (2002).


**Supplementary Information** accompanies the paper on *Nature*'s website (http://www.nature.com/nature).


## Acknowledgements

We thank our colleagues at The Wellcome Trust Sanger Institute, The Institute for Genomic Research, the Stanford Genome Technology Center, and the Naval Medical Research Center for their support. We thank J. Foster for providing markers for chromosome 14; R. Huestis and K. Fischer for providing RT–PCR data for chromosomes 2 and 3 before publication; A. Waters for assistance with ribosomal RNAs; S. Cawley for assistance with phat; and M. Crawford and R. Wang for discussions. This work was supported by the Wellcome Trust, the Burroughs Wellcome Fund, the National Institute for Allergy and Infectious Diseases, the Naval Medical Research Center, and the US Army Medical Research and Materiel Command.


## Competing interests statement

The authors declare that they have no competing financial interests.


Correspondence and requests for materials should be addressed to M.J.G. (e-mail: gardner@tigr.org).

Sequences and annotation are available at the following websites: PlasmoDB (http://plasmodb.org), The Institute for Genomic Research (http://www.tigr.org), the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/Protozoa/), and the Stanford Genome Technology Center (http://www-sequence.stanford.edu/group/malaria). Chromosome sequences were submitted to EMBL or GenBank with accession numbers AL844501–AL844509 (chromosomes 1, 3–9 and 13), AE001362.2 (chromosome 2), AE014185–AE014187 (chromosomes 10, 11 and 14) and AE014188 (chromosome 12).

# Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14

**Malcolm J. Gardner*, Shamira J. Shallom*, Jane M. Carlton*, Steven L. Salzberg*, Vishvanath Nene*, Azadeh Shoaibi*, Anne Ciecko*, Jeffery Lynn*, Michael Rizzo*, Bruce Weaver*, Behnam Jarrahi*, Michael Brenner*, Babak Parvizi*, Luke Tallon*, Azita Moazzez*, David Granger*, Claire Fujii*, Cheryl Hansen*, James Pederson†, Tamara Feldblyum*, Jeremy Peterson*, Bernard Suh*, Sam Angiuoli*, Mihaela Pertea*, Jonathan Allen*, Jeremy Selengut*, Owen White*, Leda M. Cummings*‡, Hamilton O. Smith*‡, Mark D. Adams*‡ J. Craig Venter*‡, Daniel J. Carucci†, Stephen L. Hoffman†‡ & Claire M. Fraser***

* *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
† *Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, Maryland 20910-7500, USA*

The mosquito-borne malaria parasite *Plasmodium falciparum* kills an estimated 0.7–2.7 million people every year, primarily children in sub-Saharan Africa. Without effective interventions, a variety of factors—including the spread of parasites resistant to antimalarial drugs and the increasing insecticide resistance of mosquitoes—may cause the number of malaria cases to double over the next two decades[1]. To stimulate basic research and facilitate the development of new drugs and vaccines, the genome of *Plasmodium falciparum* clone 3D7 has been sequenced using a chromosome-by-chromosome shotgun strategy[2–4]. We report

‡ Present addresses: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesdsa, Maryland 20894, USA (L.M.C.); Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA (H.O.S., M.D.A.); The Center for the Advancement of Genomics, 1901 Research Boulevard, 6th Floor, Rockville, Maryland 20850, USA (J.C.V.); Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).

**here the nucleotide sequences of chromosomes 10, 11 and 14, and a re-analysis of the chromosome 2 sequence[5]. These chromosomes represent about 35% of the 23-megabase *P. falciparum* genome.**

*P. falciparum* chromosomes were resolved on preparative pulsed field gels, and used to prepare shotgun libraries of 1–2-kilobase (kb) DNA fragments in plasmid vectors. Sequences of randomly selected clones were assembled, and gaps were closed using primer walking on plasmid templates or polymerase chain reaction (PCR) products. The cross-contamination of the chromosomal libraries with sequences from other chromosomes (up to 25%) and the high (A + T) content (80.6%) of *P. falciparum* DNA caused extreme difficulties in the gap closure process. Intergenic regions and introns frequently contained long runs of up to 50 consecutive A or T residues that were difficult to clone and sequence. The high (A + T) content of the chromosomes also prevented the construction of large insert libraries that could be used to construct scaffolds of ordered and oriented contiguous DNA sequences (contigs) during assembly. Similar but more severe problems were reported in the sequencing of the (A + T)-rich chromosome 2 of the slime mould *Dictyostelium discoideum*[6], illustrating the need to develop better methods for the cloning and sequencing of very (A + T)-rich genomes. The reported sequences contain three or four short gaps (<2 kb) in each chromosome. Contigs comprising these chromosomes were joined end-to-end before annotation. Efforts to close the remaining gaps will continue.

Examination of the sequences of chromosomes 2, 10, 11 and 14 revealed that the structure of these chromosomes was similar to that of the other chromosomes. All contained the 97–99% (A + T) putative centromeric sequences reported previously[7]. Conserved subtelomeric sequences[2] were observed in chromosomes 2, 10 and 11, but most of these elements had been deleted from both ends of chromosome 14. The termini of chromosome 14 consisted of telomeric hexamer repeats fused directly to truncated *var* (variant antigen) genes. Deletions of this type are thought to be due to chromosome breakage and healing events that occur during *in vitro* cultivation of the parasite.

Annotation procedures have improved since the publication of the *P. falciparum* chromosome 2 sequence[5]. A gene finding program, phat (pretty handy annotation tool[8]), was developed, supplementing the GlimmerM program[9] used previously. In this work, GlimmerM and phat were retrained on a larger training set of well-

Table 1 **Summary statistics**

| Feature | Value | | | | |
|---|---|---|---|---|---|
| | Whole genome | Chromosome 2 | Chromosome 10 | Chromosome 11 | Chromosome 14 |
| **The genome** | | | | | |
| Size (bp) | 22,853,764 | 947,102 | 1,694,445 | 2,035,250 | 3,291,006 |
| No. of gaps | 93 | 0 | 4 | 3 | 3 |
| Coverage* | 14.5 | 11.1 | 15.6 | 11.3 | 9.2 |
| (G + C) content (%) | 19.4 | 19.7 | 19.7 | 19.0 | 18.4 |
| No. of genes | 5,268 | 223 (209) | 403 | 492 | 769 |
| Mean gene length (bp)† | 2,283.3 | 2,079.1 (2,105.1) | 2,085.8 | 2,127.7 | 2,315.1 |
| Gene density (bp per gene) | 4,338.2 | 4,247.1 (4,531.6) | 4,204.6 | 4,136.7 | 4,279.6 |
| Percent coding | 52.6 | 49.0 (46.5) | 49.6 | 51.4 | 54.1 |
| Genes with introns (%) | 53.9 | 57.0 (43.1) | 51.4 | 50.4 | 49.9 |
| Genes with ESTs (%) | 49.1 | 46.2 | 48.1 | 48.4 | 46.9 |
| Gene products detected by proteomics‡ (%) | 51.8 | 43.5 | 49.1 | 51.0 | 52.1 |
| Exons | | | | | |
| Number | 12,674 | 510 (353) | 892 | 1,094 | 1,757 |
| Mean no. per gene | 2.4 | 2.3 (1.7) | 2.2 | 2.2 | 2.3 |
| (G + C) content (%) | 23.7 | 24.4 (24.3) | 24.5 | 23.5 | 22.8 |
| Mean length (bp) | 949.1 | 909.1 (1,246.3) | 942.3 | 956.9 | 1,013.3 |
| Total length (bp) | 12,028,350 | 463,647 (439,944) | 840,576 | 1,046,814 | 1,780,305 |
| Introns | | | | | |
| Number | 7,406 | 287 (144) | 489 | 602 | 988 |
| (G + C) content (%) | 13.5 | 13.4 (13.4) | 13.6 | 13.7 | 13.5 |
| Mean length (bp) | 178.7 | 202.4 (208.4) | 234.5 | 189.4 | 185.5 |
| Total length (bp) | 1,323,509 | 58,080 (30,006) | 114,676 | 114,012 | 183,240 |
| Intergenic regions | | | | | |
| (G + C) content (%) | 13.6 | 13.5 (14.1) | 13.6 | 14.1 | 13.2 |
| Mean length (bp) | 1,693.9 | 1,702.3 (2,063.2) | 1,678.5 | 1,768.5 | 1,717.2 |
| RNAs | | | | | |
| No. of tRNA genes | 43 | 1 | 0 | 2 | 2 |
| No. of 5S rRNA genes | 3 | 0 | 0 | 0 | 3 |
| No. of 5.8S, 18S and 28S rRNA units | 7 | 0 | 0 | 1 | 0 |
| **The proteome** | | | | | |
| Total predicted proteins | 5,268 | 223 | 403 | 492 | 769 |
| Hypothetical proteins§ | 3,208 | 121 | 265 | 339 | 485 |
| InterPro matches | 2,650 | 116 | 210 | 283 | 455 |
| Pfam matches | 1,746 | 77 | 133 | 184 | 275 |
| Gene Ontology | | | | | |
| Process | 1,301 | 63 | 89 | 110 | 168 |
| Function | 1,244 | 54 | 74 | 95 | 174 |
| Component | 2,412 | 120 | 181 | 220 | 308 |
| Targeted to apicoplast | 551 | 28 | 36 | 52 | 73 |
| Targeted to mitochondrion | 246 | 10 | 13 | 17 | 33 |
| Structural features | | | | | |
| Transmembrane domain(s) | 1,631 | 87 | 133 | 141 | 202 |
| Signal peptide | 544 | 28 | 41 | 52 | 63 |
| Signal anchor | 367 | 19 | 32 | 31 | 51 |

Numbers in parentheses under chromosome 2 indicate values obtained in the previous annotation[5]. Specialized searches used the following programs and databases: InterPro[21], Pfam[19] and Gene Ontology[22]. Predictions of apicoplast and mitochondrial targeting were performed using TargetP[26] and MitoProtII[25]; transmembrane domains, TMHMM[24]; and signal peptides and signal anchors, SignalP-2.0 (ref. 23).

*Average number of sequence reads per nucleotide. EST, expressed sequence tag.

†Excluding introns.

‡Percent of proteins detected in parasite extracts by two independent proteomic analyses[29,30].

§Hypothetical proteins are proteins with insufficient similarity to characterized proteins in other organisms to justify provision of functional assignments.

characterized genes, complementary DNAs (cDNAs) and products of PCR with reverse transcription (RT–PCR) (total length 540 kb) than was used in the earlier work. A program called Combiner was used to evaluate the GlimmerM and phat predictions, as well as the results of searches against nucleotide and protein databases, to construct consensus gene models. To assess the effect of these modifications, chromosome 2 was re-annotated and the results were compared with the previous annotation.

Application of these automated annotation procedures and manual curation of the resulting gene models for chromosome 2 produced 223 gene models. The revised procedures detected 21 genes not predicted previously, and 13 of the existing chromosome 2 models collapsed into six models in the new annotation. Of the 21 new gene models, all but one had no significant similarity to proteins in a non-redundant amino-acid database. However, at least a portion of each of the 21 gene models had been predicted independently by both GlimmerM and phat, suggesting that many of these models were likely to represent coding sequences. On the other hand, five of the new gene models encoded proteins less than 100 amino acids in length, and may be less likely to encode proteins.

Another major difference was the detection of additional small exons. In the earlier annotation of chromosome 2, the 209 predicted genes contained 353 exons, or an average of 1.7 exons per gene. The revised procedures reported here revealed 510 exons, or 2.3 exons per gene; 60% of the new exons were predicted to be additions to the gene models reported previously. Most cases involved the addition of one or two exons per gene. In three notable cases, however, 7 to 12 small exons were added to the earlier gene models, and almost all of the new exons had been predicted by both of the gene finding programs. Overall, use of the revised annotation procedures resulted in the detection of additional genes and many small exons, which is reflected in the higher gene density and shorter mean exon length in the newly annotated chromosome 2 sequence compared with the previous annotation (Table 1). Despite these improvements in software and training sets, gene finding in *P. falciparum* remains challenging, and the gene structures presented here should be regarded as preliminary until confirmed by sequence information obtained from cDNAs or RT–PCR experiments[10]. Accurate prediction of the 5′ ends of genes is particularly difficult. Generation of larger training sets, including additional expressed sequence tags (ESTs) and full-length cDNAs, would greatly improve the sensitivity and accuracy of gene predictions.

These annotation procedures were also applied to the analysis of chromosomes 10, 11 and 14 (Table 1; maps of these chromosomes are available as Supplementary Information). The 10 short gaps in the chromosomes should not have interfered with the gene predictions; only the genes adjacent to the gaps might have been affected. All three chromosomes were similar in terms of gene density, coding percentage and other parameters. A complete description of the parasite genome is contained in the accompanying Article[2].

Annotation of chromosomes 10, 11 and 14 revealed four proteins with sequence similarity to SR proteins, a family of conserved splicing factors that contain RNA-binding domains and a protein interaction domain rich in Ser and Arg residues (SR domain; PF10_0047, PF10_0217, PF11_0200, PF14_0656). Three additional putative SR proteins were identified on chromosomes 5 and 13 (PFE0160c, PFE0865c, MAL13P1.120). SR proteins are thought to bind to exonic splicing enhancers (ESEs), short (6–9 bp) sequences within exons that assist in the recognition of nearby splice sites, and to interact with components of the spliceosome[11]. ESEs have previously been characterized only in multicellular organisms. To determine whether *P. falciparum* may use ESEs as part of its splicing machinery, a Gibbs sampling algorithm for motif detection[12] was applied to a set of *P. falciparum* exons to detect any exonic splicing enhancers (ESEs). The exons were extracted from the set of well-characterized genes used to train the GlimmerM gene finder. Regions of 50 bp regions were selected from both ends of the internal exons and divided into two different data sets, representing the exon regions adjacent to both 5′ and 3′ splice sites. At least 10 runs of the Gibbs sampler were performed for each data set in order to identify the most probable motif with a length of 5–9 nucleotides. The motif with the highest maximum *a posteriori* probability was retained. This analysis identified a motif with the consensus GAAGAA, which is identical to ESEs found in human exons[13,14]. The identification of several putative SR proteins, and sequences identical to the ESEs in humans, suggests that some features of exon recognition and splicing observed in higher eukaryotes may be conserved in *P. falciparum*. □

## Methods

### Sequencing and closure

*P. falciparum* clone 3D7 was selected for sequencing because it can complete all phases of the life cycle, and had been used in a genetic cross[15] and the Wellcome Trust Malaria Genome Mapping Project[16]. High-molecular-mass genomic DNA was subjected to electrophoresis on preparative pulsed field gels, and chromosomes were excised. DNA was extracted from the gel, sheared, and cloned into the pUC18 vector as described[5] (chromosomes 2, 14) or into a modified pUC18 vector via *Bst*XI linkers (chromosomes 10, 11). Sequences were assembled and gaps were closed by primer walking on plasmid DNAs or genomic PCR products, or by transposon insertion[5]. Ordering of contigs was facilitated by the use of sequence tagged sites[16] and microsatellite markers[17]. The final assembly of each chromosome was verified by comparison with *Bam*HI and *Nhe*I optical restriction maps[18]. The average difference in size between the experimentally determined restriction fragments and the fragments predicted from the sequence was approximately 5–6% for chromosomes 11 and 14 for both enzymes. For chromosome 10, the average difference in fragment sizes was 6.1% for the *Nhe*I map, but the *Bam*HI optical and prediction restriction maps could not be aligned. Because the *Nhe*I optical restriction map agreed with that predicted from the sequence, the chromosome 10 assembly was judged to be correct.

### Annotation

GlimmerM[9] and phat[8] were trained on 117 *P. falciparum* genes and 39 cDNAs taken from GenBank, plus 32 genes from chromosomes 2 and 3 that had been verified by RT–PCR (provided by R. Huestis and K. Fischer; the training set is available at http://www.tigr.org/software/glimmerm/data). The GlimmerM and phat predictions, and sequence alignments of the chromosomes to protein and cDNA databases, were evaluated by the Combiner program. The program used a linear weighting method and dynamic programming to construct consensus gene models that were curated manually using AnnotationStation (AffyMetrix Inc.). Predicted proteins were searched against a non-redundant amino-acid database using BLASTP; other features were identified by searches against the Pfam[19], PROSITE[20] and InterPro[21] databases. The results of all analyses were reviewed using Manatee, a tool that interfaces with a relational database of the information produced by the annotation software. Predicted gene products were manually assigned Gene Ontology [22] terms. Signal peptides and signal anchors were predicted with SignalP-2.0 (ref. 23). Transmembrane helices were predicted with TMHMM[24]. Mitochondrial- and apicoplast-targeted proteins were predicted by MitoProtII[25], TargetP[26] and PATS[27]. tRNA-ScanSE[28] was used to identify transfer RNAs.

Received 6 August; accepted 2 September 2002; doi:10.1038/nature01094.


1. Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64**, 1–11 (2001).
2. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
3. Hall, N. *et al.* Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature* **419**, 527–531 (2002).
4. Hyman, R. W. *et al.* Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419**, 534–537 (2002).
5. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
6. Glockner, G. *et al.* Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418**, 79–85 (2002).
7. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
8. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118**, 167–174 (2001).
9. Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
10. Huestis, R. & Fischer, K. Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. *Mol. Biochem. Parasitol.* **118**, 187–199 (2001).
11. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243 (2002).
12. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
13. Ramchatesingh, J., Zahler, A. M., Neugebauer, K. M., Roth, M. B. & Cooper, T. A. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol. Cell Biol.* **15**, 4898–4907 (1995).
14. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).

15. Walliker, D., Quayki, I., Wellems, T. E. & McCutchan, T. F. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236**, 1661–1666 (1987).
16. Foster, J. & Thompson, J. The *Plasmodium falciparum* genome project: a resource for researchers. *Parasitol. Today* **11**, 1–4 (1995).
17. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286**, 1351–1353 (1999).
18. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23**, 309–313 (1999).
19. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
20. Falquet, L. *et al.* The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**, 235–238 (2002).
21. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
23. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6 (1997).
24. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
25. Claros, M. G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241**, 779–786 (1996).
26. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
27. Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I. & Schneider, G. Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280**, 19–26 (2001).
28. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
29. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
30. Lasonder, E. *et al.* Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542 (2002).

**Supplementary Information** accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

## Acknowledgements

We thank our colleagues at The Institute for Genomic Research and the Naval Medical Research Center for support; J. Foster for providing markers for chromosome 14; R. Huestis and K. Fischer for providing RT–PCR data for chromosomes 2 and 3 before publication; and S. Cawley for assistance with phat. This work was supported by the Burroughs Wellcome Fund, the National Institute for Allergy and Infectious Diseases, the Naval Medical Research Center, and the US Army Medical Research and Materiel Command.

## Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.G. (e-mail: gardner@tigr.org). Chromosome sequences have been deposited in GenBank with accession numbers AE001362.2 (chromosome 2), AE014185 (chromosome 10), AE01486 (chromosome 11) and AE01487 (chromosome 14), and in PlasmoDB ( http://plasmodb.org).

articles

# Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*

**Jane M. Carlton*, Samuel V. Angiuoli*, Bernard B. Suh*, Taco W. Kooij†, Mihaela Pertea*, Joana C. Silva*, Maria D. Ermolaeva*, Jonathan E. Allen*, Jeremy D. Selengut*, Hean L. Koo*, Jeremy D. Peterson*, Mihai Pop*, Daniel S. Kosack*, Martin F. Shumway*, Shelby L. Bidwell*, Shamira J. Shallom*, Susan E. van Aken*, Steven B. Riedmuller*, Tamara V. Feldblyum*, Jennifer K. Cho*‡, John Quackenbush*, Martha Sedegah§, Azadeh Shoaibi*, Leda M. Cummings*‡, Laurence Florens||, John R. Yates||, J. Dale Raine¶, Robert E. Sinden¶, Michael A. Harris#, Deirdre A. Cunningham☆, Peter R. Preiser☆, Lawrence W. Bergman**, Akhil B. Vaidya**, Leo H. van Lin†, Chris J. Janse†, Andrew P. Waters†, Hamilton O. Smith#, Owen R. White*, Steven L. Salzberg*, J. Craig Venter††, Claire M. Fraser*, Stephen L. Hoffman‡§, Malcolm J. Gardner* & Daniel J. Carucci§**

* *The Institute for Genomic Research, 9712 Medical Center Drive; and* †† *The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, Maryland 20850, USA*
† *Department of Parasitology, Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, The Netherlands*
§ *Naval Medical Research Center, Malaria Program (IDD), Silver Spring, Maryland 20910, USA*
|| *Department of Cell Biology, The Scripps Research Institute, La Jolla, California, 92037, USA*
¶ *Infection & Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, London, SW7 2AZ, UK*
# *Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA*
☆ *Division of Parasitology, National Institute for Medical Research, London, UK*
** *Division of Molecular Parasitology, Department of Microbiology & Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania 191929, USA*

**Species of malaria parasite that infect rodents have long been used as models for malaria disease research. Here we report the whole-genome shotgun sequence of one species, *Plasmodium yoelii yoelii*, and comparative studies with the genome of the human malaria parasite *Plasmodium falciparum* clone 3D7. A synteny map of 2,212 *P. y. yoelii* contiguous DNA sequences (contigs) aligned to 14 *P. falciparum* chromosomes reveals marked conservation of gene synteny within the body of each chromosome. Of about 5,300 *P. falciparum* genes, more than 3,300 *P. y. yoelii* orthologues of predominantly metabolic function were identified. Over 800 copies of a variant antigen gene located in subtelomeric regions were found. This is the first genome sequence of a model eukaryotic parasite, and it provides insight into the use of such systems in the modelling of *Plasmodium* biology and disease.**

For decades, the laboratory mouse has provided an alternative platform for infectious disease research where the pathogen under study is intractable to routine laboratory manipulation. Experimental study of the human malaria parasite *Plasmodium falciparum* is particularly problematic as the complete life cycle cannot be maintained *in vitro*. Four species of rodent malaria (*Plasmodium yoelii*, *Plasmodium berghei*, *Plasmodium chabaudi* and *Plasmodium vinckei*) isolated from wild thicket rats in Africa have been adapted to grow in laboratory rodents[1]. These species reproduce many of the biological characteristics of the human malaria parasite. Many of the experimental procedures refined for use with *P. falciparum* were initially developed for rodent malaria species, a prime example being stable genetic transformation[2]. Thus rodent models of malaria have been used widely and successfully to complement research on *P. falciparum*.

With the advent of the *P. falciparum* Genome Sequencing Project, undertaken by an international consortium of genome sequencing centres and malaria researchers, a series of initiatives has begun to generate substantial genome information from additional *Plasmodium* species[3]. We describe here the genome sequence of the rodent malaria parasite *P. y. yoelii* to fivefold genome coverage. We show that this partial genome sequencing approach, although limited in its application to the study of genome structure, has proved to be an effective means of gene discovery and of jump-starting experimental studies in a model *Plasmodium* species. Furthermore, we show that despite the considerable divergence between the *P. y. yoelii* and *P. falciparum* genomes, sequencing and annotation of the former can substantially improve the accuracy and efficiency of annotation of the latter.

## *Plasmodium yoelii yoelii* genome sequencing and annotation

We applied the whole-genome shotgun (WGS) sequencing approach, used successfully to sequence and assemble the first large eukaryotic genome[4], to achieve fivefold sequence coverage of the genome of a clone of the 17XNL line of *P. y. yoelii* (Table 1). This level of coverage is expected to comprise 99% of the genome[5] assuming random library representation. As with *P. falciparum*, the genomes of rodent malaria parasites are highly (A + T)-rich[6], which adversely affects DNA stability in plasmid libraries. Consequently, all ~220,000 reads were produced from clones originating

Table 1 ***Plasmodium yoelii yoelii* genome coverage statistics**

| Data | Component | Value |
|---|---|---|
| Genome | No. of contigs | 5,687 |
| | Mean contig size (kb) | 3.6 |
| | Max. contig size (kb) | 51.5 |
| | Cumulative contig length (Mb) | 23.1 |
| | No. of singletons | 11,732 |
| | No. of groups | 2,906 |
| | Max. group size (kb) | 69.8 |
| | Cumulative group size (Mb) | 21.6 |
| Transcriptome | No. of ESTs | 13,080 |
| | Average length (nucleotides) | 497 |
| Proteome | No. of gametocyte peptides | 1,413 |
| | No. of sporozoite peptides | 677 |

‡ Present addresses: National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA (L.M.C.); Genentech, San Francisco, California 94080, USA (J.K.C.); and Sanaria, 308 Argosy Drive, Gaithersburg, Maryland 20878, USA (S.L.H.).

from small (2–3 kilobases (kb)) insert libraries. Contigs were assembled using TIGR Assembler[7]. Contaminating mouse sequences, identified through similarity searches and found to comprise 10% of the total sequence data, were excluded from the analyses. Approximately three-quarters of the contigs could be placed into 2,906 'groups', each group consisting of two or more contigs known to be linked through paired reads as determined by Grouper software[7]. This produced an average group size of 7.4 kb, approximately 4 kb more than the average contig size. This group size is small compared with the group data produced by other partial eukaryotic genome projects, where extensive use of large insert (linking) libraries has enabled the construction of ordered and orientated 'scaffolds'[8], and emphasizes the use of such linking libraries in partial genome projects. The genome size of *P. y. yoelii* is estimated to be 23 megabases (Mb), in agreement with karyotype data[9].

Expression data from the *P. y. yoelii* transcriptome and proteome were generated to aid in gene identification and annotation of the contigs (Table 1). A total of 13,080 expressed sequence tag (EST) sequences generated from clones of an asexual blood-stage *P. y. yoelii* complementary DNA library[10], in combination with other *P. yoelii* ESTs and transcript sequences available from public databases, were assembled and used to compile a gene index[11] of expressed *P. y. yoelii* sequences (http://www.tigr.org/tdb/tgi/pygi/). For protein expression data, multidimensional protein identification technology (MudPIT), which combines high-resolution liquid chromatography with tandem mass spectrometry and database searching, was applied to the gametocyte and salivary gland sporozoite proteomes of *P. y. yoelii*. A total of 1,413 gametocyte and 677 sporozoite peptides were recorded and used for the purposes of gene annotation.

We used two gene-finding programs, GlimmerMExon and Phat[12], to predict coding regions in *P. y. yoelii*. GlimmerMExon is based on the eukaryotic gene finder GlimmerM[13], with modifications developed for analysing the short fragments of DNA that result from partial shotgun sequencing. Gene models based on GlimmerMExon and Phat predictions were refined using Combiner. Annotation of predicted gene models used TIGR's fully automated Eukaryotic Genome Control suite of programs. Gene finding and subsequent annotation were limited to 2,960 contigs (each of which is over 2 kb in size), a subset of sequences that contains more than 20 Mb of the genome. A total of 5,878 complete genes and 1,952 partial genes (defined as genes lacking either an annotated start or stop codon) can be predicted from the nuclear genome data.

Table 2 **Comparison of genome features of *P. falciparum* and *P. y. yoelii***

| Feature | *P. y. yoelii* | *P. falciparum* |
|---|---|---|
| Size (Mb) | 23.1 | 22.9 |
| No. of chromosomes | 14 | 14 |
| No. of gaps | 5,812 | 93 |
| Coverage* | 5 | 14.5 |
| (G + C) content (%) | 22.6 | 19.4 |
| No. of genes† | 5,878 | 5,268 |
| Mean gene length (bp) | 1,298 | 2,283 |
| Gene density (bp per gene) | 2,566 | 4,338 |
| Per cent coding | 50.6 | 52.6 |
| Genes with introns (%) | 54.2 | 53.9 |
| Genes with ESTs (%) | 48.9 | 49.1 |
| Gene products detected by proteomics (%) | 18.2 | 51.8 |
| Exons | | |
| Mean no. per gene | 2.0 | 2.4 |
| (G + C) content (%) | 24.8 | 23.7 |
| Mean length (bp) | 641 | 949 |
| Introns | | |
| (G + C) content (%) | 21.1 | 13.5 |
| Mean length (bp) | 209 | 179 |
| Total length (bp) | 1,687,689 | 1,323,509 |
| Intergenic regions | | |
| (G + C) content (%) | 20.7 | 13.6 |
| Mean length (bp) | 859 | 1,694 |
| RNAs | | |
| No. of tRNA genes‡ | 39 | 43 |
| No. of 5S rRNA genes | 3 | 3 |
| No. of 5.8S, 18S and 28S rRNA units | 4 | 7 |
| Mitochondrial genome | | |
| (G + C) content (%) | 31 | 31 |
| Apicoplast genome | | |
| (G + C) content (%) | 15 | 14 |

*Average number of sequence reads per nucleotide.
†Total number of full-length genes.
‡The smaller number reflect the partial nature of the *P. y. yoelii* genome data.

## Comparative genome analysis

A comparison of several genome features of *P. falciparum* and *P. y. yoelii* is shown in Table 2, demonstrating that many similarities exist between the genomes. Besides the similarly extreme (G + C) compositions, both genomes contain a comparable number of predicted full-length genes, with the higher figure in *P. y. yoelii* due to an extremely high copy number of variant antigen genes (see below). Where differences between the genomes do exist, such as the (G + C) content of the coding portion of the genomes, incompleteness of the *P. y. yoelii* genome data, with the associated problems of accurate gene finding in both species, is likely to be a confounding factor. As an indication of this problem, analysis of *P. y. yoelii* proteomic data identified 83 regions of the genome apparently expressed during sporozoite and/or gametocyte stages but not assigned to a *P. y. yoelii* gene model (data not shown). Many of these peptide hits appear sufficiently close to a model as to indicate a fault with gene boundary prediction rather than a lack of gene prediction *per se*. However, as with the gene model prediction in *P. falciparum*, the gene models of *P. y. yoelii* should be considered preliminary and under revision.

Identifying orthologues of *P. falciparum* vaccine candidate proteins and proteins that are either targets of antimalarial drugs or involved in antimalarial drug resistance mechanisms is a primary goal of model malaria parasite genomics. Using BLASTP[14] with a cutoff E value of $10^{-15}$ and no low-complexity filtering, 3,310 bidirectional orthologues (defined as genes related to each other through vertical evolutionary descent) can be identified in the full protein complement of *P. falciparum* (5,268 proteins) and the protein complement of *P. y. yoelii* translated from complete gene models (5,878 proteins). A list of vaccine candidate orthologues and orthologues of genes involved in antimalarial drug interactions identified from among the 3,310 orthologues and from additional BLAST analyses is shown in Table 3. Those genes that are not identifiable may either be absent from the partial genome data, or represent genes that have been lost or diverged sufficiently that they are undetectable through similarity searching.

Many of the candidate vaccine antigens under study in *P. falciparum* can be identified in *P. y. yoelii*, including orthologues of several asexual blood-stage antigens known to elicit immune responses in individuals exposed to natural infection (MSP1, AMA1, RAP1, RAP2). As immunity to *P. falciparum* blood-stage infection can be transferred by immune sera, identification of the targets of potentially protective antibody responses after natural infection can provide information beneficial to the selection of candidate antigens for malaria vaccines. We found several orthologues of known *P. falciparum* transmission-blocking candidates; in particular, members of the P48/45 gene family identified previously[15] were confirmed.

We identified several *P. y. yoelii* orthologues of *P. falciparum* biochemical pathway components under study as targets for drug design (Table 3), most notably: (1) the 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DOXPR) gene whose product is inhibited by fosmidomycin in *P. falciparum in vitro* cultures and mice infected with *P. vinckei*[16]; (2) enoyl-acyl carrier protein (ACP) reductase (FABI) whose product is inhibited by triclosan in *P. falciparum in vitro* cultures and mice infected with *P. berghei*[17]; and (3) a gene encoding farnesyl transferase (FTASE), which is inhibited in cultures of *P. falciparum* treated with custom-designed peptidomimetics[18]. The rodent models of malaria have proved

Table 3 ***P. y. yoelii* orthologues of *P. falciparum* candidate vaccine and drug interaction genes**

| *P. falciparum* gene | *Pf* chromosome | ST location* | *Pf* locus | *Py* locus |
|---|---|---|---|---|
| **Candidate vaccine antigens** | | | | |
| Ring-infected erythrocytic surface antigen 1, *resa1* | 1 | Yes | PFA0110w | Not identified |
| Merozoite surface protein 4, *msp4* | 2 | No | PFB0310c | PY07543† |
| Merozoite surface protein 5, *msp5* | 2 | No | PFB0305c | PY07543† |
| Liver stage antigen 3, *lsa3* | 2 | No | PFB0915w | Not identified |
| Merozoite surface protein 2, *lsa3* | 2 | No | PFB0300c | Not identified |
| Transmission-blocking target antigen 230, *Pfs230* | 2 | No | PFB0405w | PY03856 |
| Circumsporozoite protein, *csp* | 3 | No | MAL3P2.11 | PY03168 |
| Rhoptry-associated protein 2, *rap2* | 5 | Yes | PFE0080c | PY03918 |
| Sporozoite surface antigen, *starp* | 7 | Yes | PF07_0006 | Not identified |
| Morozoite surface protein 1, *msp1* | 9 | No | PFI475w | PY05748 |
| Liver stage antigen 1, *lsa1* | 10 | No | PF10_0356 | Not identified |
| Merozoite surface protein 3, *msp3* | 10 | No | PF10_0345 | Not identified |
| Glutamate-rich protein, *glurp* | 10 | No | PF10_0344 | Not identified |
| Ookinete surface protein 25, *Pfs25* | 10 | No | PF10_0303 | PY00523 |
| Ookinete surface protein 28, *Pfs28* | 10 | No | PF10_0302 | PY00522 |
| Erythrocyte membrane-associated 332 antigen, *Pf332* | 11 | No | PF11_0507 | PY06496 |
| Apical membrane antigen 1, *ama1* | 11 | No | PF11_0344 | PY01581 |
| Exported protein 1, *exp1* | 11 | No | PF11_0224 | Not identified |
| Surface sporozoite protein 2, *ssp2* | 13 | No | PF13_0201 | PY03052 |
| Sexual-stage-specific surface antigen 48/45, *Pfs48/45* | 13 | No | PF13_0247 | PY04207 |
| Rhoptry-associated protein 1, *rap1* | 14 | Yes | PF14_0637 | PY00622 |
| **Candidate drug interaction genes** | | | | |
| Dihydrofolate reductase, *dhfr* | 4 | No | PFD0830w | PY04370 |
| Multidrug resistance protein 1, *pfmdr1* | 5 | No | PFE1150w | PY00245 |
| Translationally controlled tumour protein, *tctp* | 5 | No | PFE0545c | PY04896 |
| Farnesyl transferase, *ftase* | 5 | No | PFE0970w | PY06214 |
| Enoyl-acyl carrier reductase, *fabi* | 6 | No | MAL6P1.275 | PY03846 |
| Dihydro-protate dehydrogenase, *dhod* | 6 | No | MAL6P1.36 | PY02580 |
| Chloroquine-resistance transporter, *pfcrt* | 7 | No | MAL7P1.27 | PY05061 |
| Dihydropteroate synthase, *dhps* | 8 | No | PF08_0095 | PY02226 |
| Lactate dehydrogenase, *ldh* | 13 | No | PF13_0141 | PY03885 |
| DOXP reductoisomerase, *doxpr* | 14 | No | PF14_0641 | PY05578 |

A full listing of all orthologues can be found as Table A in the Supplementary Information. *Pf*, *P. falciparum*; *Py*, *P. y. yoelii*.
*ST, subtelomeric. Defined as >75% of the distance from the centre to the end of the *P. falciparum* chromosome.
†Homologue of *P. falciparum msp4* and *msp5* genes found as a single gene *msp4/5* in *P. y. yoelii* and other rodent malaria species[62].

invaluable both for the study of potency of new antimalarial compounds *in vivo*, and for the elucidation of mechanisms of antimalarial drug resistance.

We applied the Gene Ontology (GO) gene classification system[19], which uses a controlled vocabulary to describe genes and their function, to indicate which classes of gene among the 3,310 orthologues might differ in number between *P. falciparum* and *P. y. yoelii* (Fig. 1). A similar proportion of proteins were identified for most of the GO classes between the two species, with the caveat that fewer total numbers of proteins were identified in *P. y. yoelii* owing to the partial nature of the genome data for this species. However, proteins allocated to the physiological processes, cell invasion and adhesion, and cell communication categories were significantly reduced in *P. y. yoelii*. These classes contain members of three multigene families whose genes are found predominantly in the subtelomeric regions of *P. falciparum* chromosomes: PfEMP1, the protein product of the *var* gene family known to be involved in antigenic variation, cyto-adherence and rosetting, and rifins and stevors, which are clonally variant proteins possibly involved in antigenic variation and evasion of immune responses (reviewed in ref. 20). Apparently, *P. falciparum* has generated species-specific, subtelomeric genes involved in host cell invasion, adhesion and antigenic variation, homologues of which are not found in the *P. y. yoelii* genome.

## Gene families of unique interest in the *P. y. yoelii* genome

The largest family of genes identified in the *P. y. yoelii* genome is the *yir* gene family, homologues of the *vir* multigene family recently described in the human malaria parasite *Plasmodium vivax*[21] and in other species of rodent malaria[22]. In *P. vivax*, an estimated 600–1,000 copies of the subtelomerically located *vir* gene encode proteins that are immunovariant in natural infections, indicating a possible functional role in antigenic variation and immune evasion. Within the *P. y. yoelii* genome data, 838 *yir* genes (693 full genes and 145 partial genes) are present (Table 4; see also Supplementary Figs A and B). Almost 75% of the annotated contigs identified as containing subtelomeric sequences (see below) contain *yir* genes, many arranged in a head-to-tail fashion. Expression data indicate that *yir* genes are expressed during sporozoite, gametocyte and erythrocytic stages of the parasite, similar to the expression pattern seen with *P. falciparum var* and *rif* genes[23]. Preliminary

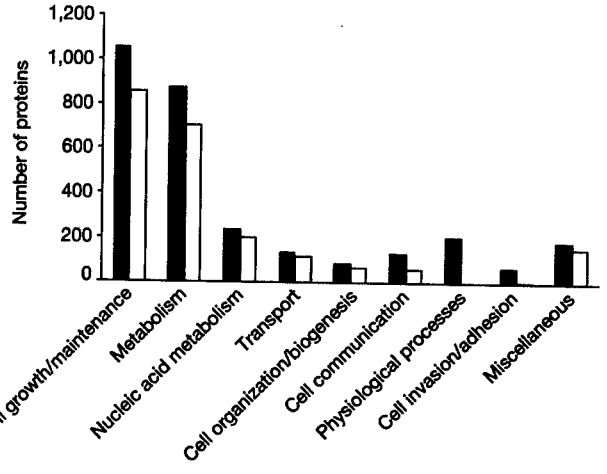



**Figure 1** Functional classification comparison between *P. falciparum* and *P. y. yoelii* proteins. We compared the GO terms of proteins assigned to 'biological process' for the orthologous genes identified between the two species. The process group contains 3,041 *P. falciparum* annotations (filled bars), and 2,161 reciprocal annotations are shown for *P. y. yoelii* (open bars). Ten GO classes with similar numbers of *P. falciparum* and *P. y. yoelii* proteins in each are assigned as 'miscellaneous'; that is, cell cycle, external stimulus response, stress response, signal transduction, homeostasis, developmental processes, cell proliferation, membrane fusion, death, cell motility.

Table 4 **Paralogous gene families in *P. y. yoelii***

| Gene family | No. | Name | HMM ID | Location in *Py* | *Py* expression* | *Pf* locus | TM/SP† |
|---|---|---|---|---|---|---|---|
| *yir/bir/cir* | 838 | Variant antigen family | TIGR01590 | Subtelomeric | Gmt, spz, bs | None | P/A |
| *235 kDa* | 14 | Reticulocyte binding family | TIGR01612 | Subtelomeric | Gmt, spz, bs | PFD0110w, MAL13P1.176, PF13_0198, PFL2520w, PFD0110w | P/A |
| *pyst-a* | 168 | Hypothetical | TIGR01599 | Subtelomeric | Gmt, spz | PF14_0604 | A/A |
| *pyst-b* | 57 | Hypothetical | TIGR01597 | Subtelomeric | Bs | None | P/A |
| *pyst-c* | 21 | Hypothetical | TIGR01601, TIGR01604 | Subtelomeric | Bs | None | P/P |
| *pyst-d* | 17 | Hypothetical | TIGR01605 | Subtelomeric | Gmt | None | P/P |
| *etramp* | 11 | Early transcribed membrane protein family | TIGR01495 | Subtelomeric | Gmt, spz, bs | PF13_0012, PF14_0016, PF11_0040, PFB0120w, PF10_0323, MAL12P1.387, PF11_0039, PFL1095c, PF10_0019, PF1745c, PFE1590w, PF10_0164, MAL8P1.6, PFA0195w, PFL0065w, PF14_0729 | P/P |
| *pst-a* | 12 | Hydrolase family | TIGR01607 | Subtelomeric | Gmt, spz | PFL2530w, PF10_0379, PF14_0738, PF14_0017, PF14_0737, PFI800w, PFI1775w, PF07_0040, PF07_0005, PFA0120c | A/A |
| *rhoph1/clag* | 2 | Rhoptry H1/ cyto-adherence-linked asexual gene family | PF03805 | Subtelomeric | Gmt, bs | PFC0110w, PFC0120w, PFI1730w, PFI1710w, PFB0935w | A/P |

*Found in, but not limited to: gmt, gametocyte life stage; spz, sporozoite life stage; bs, asexual blood stage.
†TM, transmembrane domain; SP, signal peptide; P, predicted; A, absent. TM and SP predictions were identical for *P. falciparum* and *P. y. yoelii* members of the same gene family. (See ref. 30 for details regarding TM and SP prediction algorithms.)

results using antibodies developed against the conserved regions of the protein have confirmed protein localization at the surface of the infected red blood cell (D.A.C. *et al.*, manuscript in preparation). The number of gene copies in the *P. y. yoelii* genome, the localization and stage-specific expression of gene members, as well as the existence of homologues in other *Plasmodium* species, make this gene family a prime target for the study of mechanisms of immune evasion.

A maximum of 14 members of the *Py235* multigene family can be identified among the *P. y. yoelii* protein data (Table 4). This family expresses proteins that localize to rhoptries (organelles that contain proteins involved in parasite recognition and invasion of host red blood cells). *Py235* genes exhibit a newly discovered form of clonal antigenic variation, whereby each individual merozoite derived from a single parent schizont has the propensity to express a different Py235 protein[24]. Closely related homologues of the *Py235* gene family have been found in other rodent malaria species, and more distantly related homologues have been found in *P. vivax*[25] and *P. falciparum*[26]. The gene copy number identified in the current data set is less than has been predicted in other *P. y. yoelii* lines (30–50 per genome). This could reflect real differences in copy number between lines, but more probably suggests an error in the original estimate or misassembly of extremely closely related sequences. Almost all of the *Py235* genes are found on contigs identified as subtelomeric in the *P. y. yoelii* genome (see Supplementary Fig. C).

Four further paralogous gene families, *pyst-a* to *-d*, are specific to *P. y. yoelii* (Table 4). The *pyst-a* family deserves mention, as it is homologous to a *P. chabaudi* glutamate-rich protein[27] and to a single hypothetical gene on *P. falciparum* chromosome 14, suggesting expansion of this family in the rodent malaria species from a common ancestral *Plasmodium* gene. Two paralogous gene families containing multiple members are homologous to multigene families identified in *P. falciparum*. Gene members of one family, *etramp* (early transcribed membrane protein), have previously been identified in *P. falciparum*[28] and in *P. chabaudi* where a single member has been identified and localized to the parasitophorous vacuole membrane[29].

## Telomeres and chromosomal exchange in subtelomeric regions

The telomeric repeat in *P. y. yoelii* is AACCCTG, which differs from the *P. falciparum* telomeric repeat AACCCTA by one nucleotide. A total of 71 contigs were found to contain telomeric repeat sequences arranged in tandem, with the largest array consisting of 186 copies. The *P. y. yoelii* subtelomeric chromosomal regions show little repeat structure compared with those of *P. falciparum*. A survey of tandem repeats in the entire genome found only a few in the telomeric or subtelomeric regions, specifically a 15 base pair (bp) (45 copies) and a 31-bp (up to 10 copies), both of which were found on multiple contigs, and a 36-bp repeat that occurred on one contig. No repeat element that corresponds to Rep20, a highly variable 21-bp unit that spans up to 22 kb in *P. falciparum* telomeres, was found.

The telomeric and subtelomeric regions of *P. y. yoelii* contigs show extensive large-scale similarity, indicating that these regions undergo chromosomal exchange similar to that reported for *P. falciparum* (see ref. 30). The longest subtelomeric contig is approximately 27 kb (see Supplementary Fig. C) and is homologous to other subtelomeric contigs across its entire length, indicating that the region of chromosomal exchange extends at least this distance into the subtelomeres. Recent data have shown that clustering of telomeres at the nuclear periphery in asexual and sexual stage *P. falciparum* parasites may promote sequence exchange between members of subtelomeric virulence genes on heterologous chromosomes, resulting in diversification of antigenic and adhesive phenotypes (see ref. 31 for review). The suggestion of extensive chromosome exchange in *P. y. yoelii* indicates that a similar system for generating antigenic diversity of the *yir*, *Py235* and other gene families located within subtelomeric regions may exist.

## A genome-wide synteny map

The *Plasmodium* lineage is estimated to have arisen some 100–180 million years ago[32], and species of the parasite are known to infect birds, mammals and reptiles[33]. On the basis of the analysis of small subunit (SSU) ribosomal RNA sequences, the closest relative to *P. falciparum* is *Plasmodium reichenowi*, a parasite of chimpanzees, with the rodent malaria species forming a distinct clade[34,35]. Early gene mapping studies have shown that regions of gene synteny exist between species of rodent malaria[9] and between human malaria species[36,37], despite extensive chromosome size polymorphisms between homologous chromosomes[38]. This level of gene synteny seems to decrease as the phylogenetic distance between *Plasmodium* species increases[39]. Before the *Plasmodium* genome sequencing

projects, the degree to which conservation of synteny extended across *Plasmodium* genomes was not fully apparent.

Using the *P. falciparum* and *P. y. yoelii* genome data, we have constructed a genome-wide syntenic map between the species. To avoid confounding factors inherent in DNA-based analyses of (A + T)-rich genomes, we first calculated the protein similarity between all possible protein-coding regions in both data sets using MUMmer[40]. Sensitivity was ensured through the use of a minimum word match length of five amino acids chosen to identify seed maximal unique matches (MUMs). By comparison, the recent human–mouse synteny analysis used a match length of 11 (ref. 8). Using this method, which is independent of gene prediction data, 2,212 sequences could be aligned (tiled) to *P. falciparum* chromosomes, representing a cumulative length of 16.4 Mb of sequence, or over 70% of the *P. y. yoelii* genome (see Supplementary Table C). The per cent of each *P. falciparum* chromosome covered with *P. y. yoelii* matches varies from 12% (chromosome 4) to 22% (chromosomes 1 and 14), with an average of about 18%. The spatial arrangement of the tiling paths (see Fig. 1 in ref. 30) confirms previous suggestions[9] that most of the conserved matches are found within the body of *Plasmodium* chromosomes, and confirms the absence of *var*, *rif* and *stevor* homologues in the *P. y. yoelii* genome.

Although the tiling paths indicate the degree of conservation of gene order between *P. falciparum* and *P. y. yoelii*, longer stretches of contiguous *P. y. yoelii* sequence are necessary to examine this feature in depth. Accordingly, we carried out linkage of many *P. y. yoelii* assemblies adjacent to each other along the tiling paths. First, 1,050 adjacent contigs were linked on the basis of paired reads as determined by Grouper software. Second, *P. y. yoelii* ESTs were aligned to the tiling paths, and those found to overlap sequences adjacent in the tiling path were used as evidence to link a further 236 *P. y. yoelii* sequences. Third, amplification of the sequence between adjacent contigs in the tiling paths linked a further 817 assemblies. Linkage of *P. y. yoelii* sequences by these methods resulted in the formation of 457 syntenic groups from 2,212 original contigs, ranging in length from a few kilobases to more than 800 kb. Syntenic groups were assigned to a *P. y. yoelii* chromosome where possible through the use of a partial physical map[9]. Thus, long contiguous sections of the *P. y. yoelii* genome with accompanying *P. y. yoelii* chromosomal location can be assigned to each *P. falciparum* chromosome (see Fig. 1 in ref. 30). The degree of conservation of gene order between the species was examined using ordered and orientated syntenic groups and Position Effect software. Of 4,300 *P. y. yoelii* genes within the syntenic groups, 3,145 (73%) were found to match a region of *P. falciparum* in conserved order.

One section of the syntenic map between *P. falciparum* and *P. y. yoelii* in particular—associated with *P. falciparum* chromosomes 4 and 10 and *P. y. yoelii* chromosome 5—provides a detailed snapshot of synteny between the species. Chromosome 5 of *P. y. yoelii* has received particular attention owing to the localization of a number of sexual-stage-specific genes to it[41], and because truncated versions of the chromosome are found in lines of the rodent malaria parasite *P. berghei*, which is defective in gametocytogenesis[42]. Genomic resources available for *P. berghei* chromosome 5 include chromosome markers and long-range restriction maps[41]. Exploiting the high level of synteny of rodent malaria parasite chromosomes[9], these tools were applied in combination with further mapping studies to close the syntenic map of chromosome 5 of *P. y. yoelii* (Fig. 2).

Approximately 0.8 Mb of *P. y. yoelii* chromosome 5 (estimated total length of 1.5 Mb) could be linked into one group that is syntenic to *P. falciparum* chromosome 10 and *P. falciparum* chromosome 4. From a total of 243 genes predicted in the syntenic region of *P. falciparum* chromosome 10, and 34 genes predicted in the syntenic region of chromosome 4, 171 (70%) and 22 (65%) of these, respectively, have homologues along *P. y. yoelii* chromosome 5 that appear in the same order. Pairs of homologous genes that map to regions of conserved synteny between *P. y. yoelii* and *P. falciparum* are probably orthologues, confirmed by the finding that most of these homologous pairs are also reciprocal best matches between the *P. falciparum* and *P. y. yoelii* proteins. Genes in the synteny gap on chromosome 10 (Fig. 2) include a glutamate-rich protein, S antigen, MSP3, MSP6 and liver stage antigen 1, several of which are prime vaccine antigen candidates in *P. falciparum*. Genes in the synteny gap on chromosome 4 include four *var* and two *rif* genes, which make up one of the four internal clusters of *var/rif* genes found in *P. falciparum* (see ref. 30). A series of uncharacterized hypothetical genes occur on the contigs that overlap these regions in *P. y. yoelii*.

An intriguing finding from the study of chromosome 5 has been the analysis of the syntenic break point between *P. falciparum* chromosomes 4 and 10. The final *P. y. yoelii* contig in the tiling path with significant synteny to *P. falciparum* chromosome 10 also contains the external transcribed sequence (ETS) of the SSU rRNA C unit. The synteny resumes on *P. falciparum* chromosome 4 in a *P. y. yoelii* contig that also contains the ETS of the large subunit (LSU) of the same rRNA unit. (No rRNA unit sequences are located on *P. falciparum* chromosomes 4 and 10; matches to contigs containing these genes occur in coding regions of other genes.) Both *P. y. yoelii* contigs are linked to each other through a third contig that contains the remaining elements (SSU, 5.8S, LSU, and internal transcribed sequences 1 and 2) of the complete rRNA unit (Fig. 2). Thus it seems that the break in synteny between *Plasmo-*

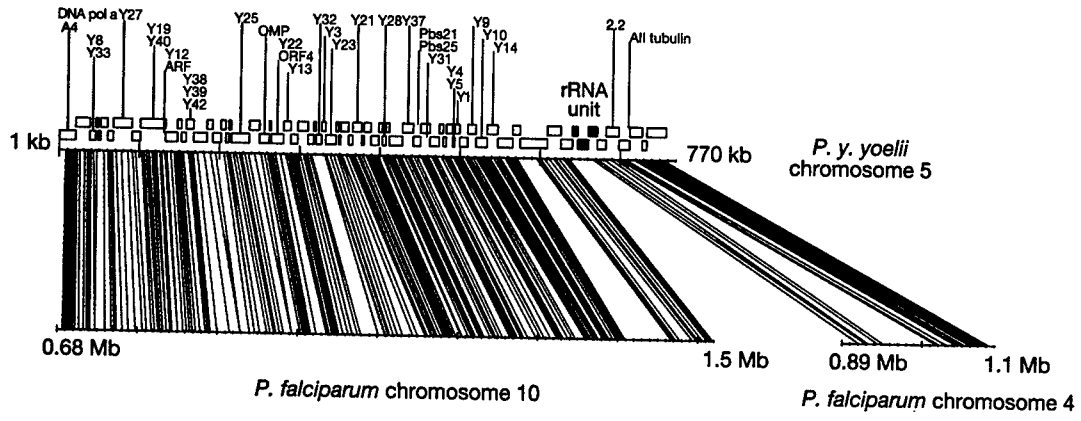



**Figure 2** Conservation of gene synteny between *P. y. yoelii* chromosome 5 and *P. falciparum* chromosomes 4 and 10. Physical marker data used to confirm contig order in the tiling path of *P. y. yoelii* chromosome 5 are shown above the contigs (open boxes). Each coloured line represents a pair of orthologous genes present in the two species shown anchored to its respective location in the two genomes. Contigs containing the *P. y. yoelii* rRNA unit are shown as filled boxes.

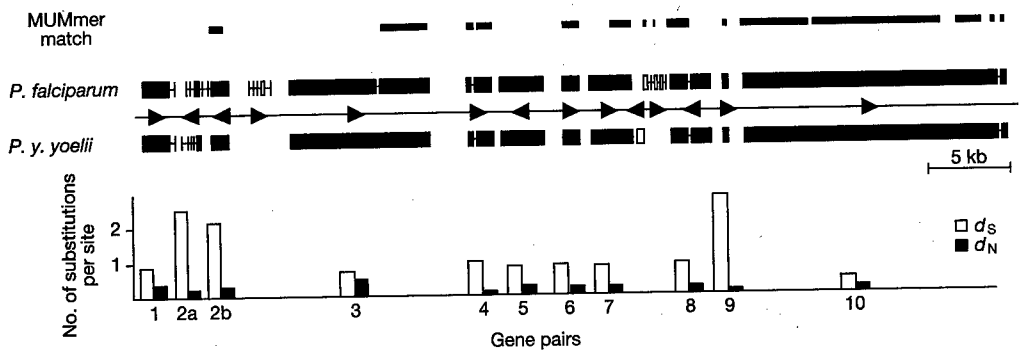



**Figure 3** Global alignment scheme of a syntenic region between *P. falciparum* and *P. y. yoelii* encompassing ten orthologous gene pairs and nine intergenic regions. White boxes represent genes that have no orthologue and were excluded from analysis; green boxes represent gene models that were refined; red boxes represent unaltered gene models; arrowheads represent gene orientation on the DNA molecule. Clusters of MUMmer matches between the two species are represented as thick blue lines. For the ten orthologous gene pairs, synonymous mutations per synonymous site ($d_S$, open bars) and non-synonymous mutations per non-synonymous site ($d_N$, filled bars) were estimated and plotted.

*dium* chromosomes has occurred within a single rRNA unit, a phenomenon first reported in prokaryotes[43]. Six rRNA units reside as individual operons on *P. falciparum* chromosomes 1, 5, 7, 8, 11 and 13 respectively (ref. 30), in contrast to rodent malaria species that have four[44]. Intriguingly, breaks in the synteny between *P. y. yoelii* and *P. falciparum* can be mapped to almost all rRNA unit loci on the *P. falciparum* chromosomes (see Fig. 1 of ref. 30). A full analysis of this potential phenomenon is outside the scope of this study, but these results provide preliminary evidence for one possible mechanism underlying synteny breakage that may have occurred during evolution of the *Plasmodium* genus—that of chromosome breakage and recombination at sites of rRNA units.

## Comparative alignment of syntenic regions

Recent comparative studies have revealed that the fine detail of short stretches of the rodent and human malaria parasite genomes is remarkably conserved[45], and that such comparisons are useful for gene prediction and evolutionary studies. Accordingly, we used a comparison of the longest assembly of *P. y. yoelii* (MALPY00395, 51.3 kb) and its syntenic region in *P. falciparum* (chromosome 7, at coordinates 1,131–1,183 kb) as a case study for a preliminary evolutionary analysis of the two genomes. Gene prediction programs run against these two regions identified 11 genes in the syntenic region of both species (Fig. 3), eight of which are orthologous gene pairs (genes 1, 3–8 and 10). The structures of two additional gene pairs (genes 2a/b and 9) were refined through manual curation of erroneous gene boundaries. Three hypothetical genes, two in *P. falciparum* and one in *P. y. yoelii*, had no discernible orthologue in the other species; the presence of multiple stop codons in these areas suggests that the genes may have become pseudogenes. A global alignment at the DNA level of the syntenic region (Fig. 3) reveals the similarity between species in intergenic regions to be almost negligible, as mirrored in similar syntenic comparisons of mouse and human[46,47]. Moreover, the mutation saturation observed in intergenic regions suggests that 'phylogenetic footprinting' can be used to identify conserved motifs between species that may be involved in gene regulation.

In contrast to intergenic regions, the similarity between species in coding regions is relatively high. The average number of non-synonymous substitutions per non-synonymous site, $d_N$, between the two species is 26% (±12%). Synonymous sites, $d_S$, are saturated (average $d_S > 1$), which supports the lack of similarity observed within intergenic regions. These values are considerably higher than those reported for human–rodent comparisons, which are approximately 7.5% and 45% for non-synonymous and synonymous substitutions, respectively[48]. The cause of such apparent disparities remains unknown, but may be a consequence of extreme genome composition or the short generation time of the parasite.

## Rodent malaria species as models for *P. falciparum* biology

The usefulness of rodent malaria species as models for the study of *P. falciparum* is controversial. It is apparent that rodent models are the first port of call when preliminary *in vivo* evidence of antimalarial drug efficacy, immune response to vaccine candidates, and life-cycle adaptations in the face of drug or vaccine challenge are required. Different species of malaria parasite have developed different mechanisms of resistance to the antimalarial drug chloroquine, despite a similar mode of action of the drug (reviewed in ref. 49). It seems that mechanisms developed by the parasite to evade an inhospitable environment, whether caused by antimalarial drugs or the host immune system, may differ widely from species to species. A model involving evolution of different genes in *Plasmodium* species as a response to different host environments is consistent with the comparison of the *P. falciparum* and *P. y. yoelii* genomes presented here; conservation of synteny between the two species is high in regions of housekeeping genes, but not in regions where genes involved in antigenic variation and evasion of the host immune system are located. On the one hand, this can be interpreted as a blow to the systematic identification of all orthologues of antigen genes between *P. falciparum* and *P. y. yoelii* that could be used in the design of a malaria vaccine. On the other hand, a picture is emerging of selecting a model malaria species based on the complement of genes that best fit the phenotypic trait under study. Thus the presence of homologues of the *yir* family may make *P. y. yoelii* an attractive model for studying antigenic variation in *P. vivax*. Furthermore, identification of orthologues in the genomes of relatively distant rodent and human malaria parasites will facilitate finding orthologues in other model malaria species, for example monkey models of malaria such as *Plasmodium knowlesi*. □

## Methods

### Genome and EST sequencing

*Plasmodium yoelii yoelii* 17XNL line[50], selected from an isolate taken from the blood of a wild-caught thicket rat in the Central African Republic[51], is a non-lethal strain with a preference for development in reticulocytes. Clone 1.1 was obtained through serial dilution of sporozoites. Parasites were grown in laboratory mice no more than three blood passages from mosquito passage to limit chromosome instability, collected by exsanguination into heparin, and host mouse leukocytes were removed by filtration. Small insert libraries (average insert size 1.6 kb) were constructed in pUC-derived vectors after nebulization of genomic DNA. DNA sequencing of plasmid ends used ABI Big Dye terminator chemistry on ABI3700 sequencing machines. A total of 222,716 sequences (82% success rate), averaging 662 nucleotides in length, were assembled using TIGR Assembler[7]. BLASTN of the *P. y. yoelii* contigs and singletons against the complete set of

Celera mouse contigs[8], using a cutoff of 90% identity over 100 nucleotides, identified contaminating mouse sequences that were subsequently removed. Contigs were assigned to groups using Grouper[52]. Each contig was assigned an identifier in the format 'MALPY00001'.

### Proteomic analysis

MudPIT technology and methods were as described in ref. 23. Sporozoites of *P. y. yoelii* were dissected from infected *Anopheles stephensi* mosquito salivary glands, and *P. y. yoelii* gametocytes were prepared as described[53]. Cellular debris from uninfected mosquitoes and mouse erythrocytes were analysed as controls. Tandem mass spectrometry (MS/MS) data sets were searched against several databases: the complete set of *P. y. yoelii* full and partial proteins (7,860 total); 791,324 *P. y. yoelii* open reading frames (stop-to-stop ORFs over 15 amino acids and start-to-stop ORFs over 100 amino acids); 57,885 ORFs from NCBI's RefSeq for human, mouse and rat; 15,570 *Anopheles*, *Aedes* and *Drosophila melanogaster* proteins from GenBank; and 165 common protein contaminants (for example, trypsin, bovine serum albumin).

### Gene finding and annotation

The splice site recognition module of GlimmerMExon was trained specifically for *P. yoelii* genome data, using DNA sequences extracted from a set of 1,166 donor and 1,166 acceptor sites confirmed by *P. y. yoelii* ESTs. Phat and the exon recognition module of GlimmerMExon were trained on *P. falciparum* data as described (see ref. 54). Combiner was used to generate a final ranked list of *P. y. yoelii* gene models, and TIGR's Eukaryotic Genome Control suite of programs was used for automated annotation of these (both described in ref. 54). Automated gene names were assigned to proteins by taking the 'equivalogue' name of the hidden Markov model (HMM) associated with the protein where possible, or where no HMM was assigned, on the basis of the best-paired alignment. Each protein was assigned an identifier in the format 'PY00001'.

### Paralogous gene families

Proteins encoded by multigene families were identified by a domain-based clustering algorithm developed at TIGR. Families were regarded as potentially *Plasmodium*- or *yoelii*-specific if they were not described by any Pfam[55] or TIGRFAM[56] domains and if the automatic annotation process had not ascribed names corresponding to widely distributed proteins. HMMs for these families were built using the HMMER package version 2.1.1 (ref. 57). Newly constructed models were then used to search the *P. yoelii*, *P. falciparum* and GenBank databases to define the scope of the families.

### Telomeric/subtelomeric repeat analysis

Subtelomeric contigs were identified through alignment using MUMmer2 (ref. 40) with a minimum exact match ranging from 30–40 bases. Tandem Repeat Finder[58] used the following settings: match = 2, mismatch = 7, PM (match probability) = 75, PI (indel probability) = 10, minscore = 400, max period = 700.

### Comparative analyses

Gene model predictions in the syntenic region of *P. falciparum* chromosome 7 were inspected manually, and bi-directional best hits between gene models that respected conserved syntenies were selected. A global alignment of the two sequences was calculated using Owen[59], and nucleotide sequences of predicted gene models were aligned using CLUSTALW[60] with default parameters, and refined manually. The number of substitutions per synonymous ($d_S$) and nonsynonymous ($d_N$) sites were estimated using the Nei and Gojobori method[61]. Conservation of gene order was established using Position Effect (http://www.tigr.org/software), where matches between *P. falciparum* and *P. y. yoelii* genes were calculated using BLASTP with a cutoff E value of $10^{-15}$. The query and hit gene from each match were defined as anchor points in gene sets composed of adjacent genes. Up to ten genes upstream and downstream from each anchor gene were used in creating the gene set. An optimal alignment was calculated between the ordered gene sets using BLASTP per cent similarity scores and a linear gap penalty. Low-scoring alignments with a cumulative per cent similarity less than 100 were not used. Each optimal alignment provided a list of matching genes in conserved order between *P. falciparum* and *P. y. yoelii*.

Received 31 July; accepted 30 August 2002; doi:10.1038/nature01099.


1. Carter, R. & Diggs, C. L. *Parasitic Protozoa* 359–465 (Academic, New York/San Francisco/London, 1977).
2. van Dijk, M. R., Waters, A. P. & Janse, C. J. Stable transfection of malaria parasite blood stages. *Science* **268**, 1358–1362 (1995).
3. Carlton, J. M. & Carucci, D. J. Rodent models of malaria in the genomics era. *Trends Parasitol.* **18**, 100–102 (2002).
4. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
5. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
6. McCutchan, T. F., Dame, J. B., Miller, L. H. & Barnwell, J. Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA. *Science* **225**, 808–811 (1984).
7. Sutton, G. G., White, O., Adams, M. D. & Kervalage, A. R. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1**, 9–19 (1995).
8. Mural, R. J. *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671 (2002).
9. Janse, C. J., Carlton, J. M.-R., Walliker, D. & Waters, A. P. Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. *Mol. Biochem. Parasitol.* **68**, 285–296 (1994).
10. Daly, T. M., Long, C. A. & Bergman, L. W. Interaction between two domains of the *P. yoelii* MSP-1 protein detected using the yeast two-hybrid system. *Mol. Biochem. Parasitol.* **117**, 27–35 (2001).
11. Quackenbush, J. *et al.* The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159–164 (2001).
12. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118**, 167–174 (2001).
13. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
15. Thompson, J., Janse, C. J. & Waters, A. P. Comparative genomics in *Plasmodium*: a tool for the identification of genes and functional analysis. *Mol. Biochem. Parasitol.* **118**, 147–154 (2001).
16. Jomaa, H. *et al.* Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**, 1573–1576 (1999).
17. Surolia, N. & Surolia, A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. *Nature Med.* **7**, 167–173 (2001).
18. Ohkanda, J. *et al.* Peptidomimetic inhibitors of protein farnesyltransferase show potent antimalarial activity. *Bioorg. Med. Chem. Lett.* **11**, 761–764 (2001).
19. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
20. Cooke, B. M., Mohandas, N. & Coppel, R. L. The malaria-infected red blood cell: structural and functional changes. *Adv. Parasitol.* **50**, 1–86 (2001).
21. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410**, 839–842 (2001).
22. Janssen, C. S., Barrett, M. P., Turner, C. M. & Phillips, R. S. A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proc. R. Soc. Lond. B* **269**, 431–436 (2002).
23. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
24. Preiser, P. R., Jarra, W., Capiod, T. & Snounou, G. A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria. *Nature* **398**, 618–622 (1999).
25. Galinski, M. R., Xu, M. & Barnwell, J. W. *Plasmodium vivax* reticulocyte binding protein-2 (PvRBP-2) shares structural features with PvRBP-1 and the *Plasmodium yoelii* 235 kDa rhoptry protein family. *Mol. Biochem. Parasitol.* **108**, 257–262 (2000).
26. Rayner, J. C., Galinski, M. R., Ingravallo, P. & Barnwell, J. W. Two *Plasmodium falciparum* genes express merozoite proteins that are related to *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins involved in host cell selection and invasion. *Proc. Natl Acad. Sci. USA* **97**, 9648–9653 (2000).
27. Wiser, M. F., Giraldo, L. E., Schmitt-Wrede, H. P. & Wunderlich, F. *Plasmodium chabaudi*: immunogenicity of a highly antigenic glutamate-rich protein. *Exp. Parasitol.* **85**, 43–54 (1997).
28. Spielmann, T. & Beck, H. P. Analysis of stage-specific transcription in *Plasmodium falciparum* reveals a set of genes exclusively transcribed in ring stage parasites. *Mol. Biochem. Parasitol.* **111**, 453–458 (2000).
29. Favaloro, J. M., Culvenor, J. G., Anders, R. F. & Kemp, D. J. A *Plasmodium chabaudi* antigen located in the parasitophorous vacuole membrane. *Mol. Biochem. Parasitol.* **62**, 263–270 (1993).
30. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
31. Scherf, A., Figueiredo, L. M. & Freitas-Junior, L. H. *Plasmodium* telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.* **4**, 409–414 (2001).
32. Mu, J. *et al.* Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418**, 323–326 (2002).
33. Garnham, P. C. C. *Malaria Parasites and Other Haemosporidia* (Blackwell Scientific, Oxford, 1966).
34. Escalante, A. A. & Ayala, F. J. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc. Natl Acad. Sci. USA* **91**, 11373–11377 (1994).
35. Waters, A. P., Higgins, D. G. & McCutchan, T. F. *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc. Natl Acad. Sci. USA* **88**, 3140–3144 (1991).
36. Tchavtchitch, M., Fischer, K., Huestis, R. & Saul, A. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. *Mol. Biochem. Parasitol.* **118**, 211–222 (2001).
37. Carlton, J. M.-R., Galinski, M. R., Barnwell, J. W. & Dame, J. B. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Mol. Biochem. Parasitol.* **101**, 23–32 (1999).
38. Janse, C. J. Chromosome size polymorphism and DNA rearrangements in *Plasmodium*. *Parasitol. Today* **9**, 19–22 (1993).
39. Carlton, J. M. R., Vinkenoog, R., Waters, A. P. & Walliker, D. Gene synteny in species of *Plasmodium*. *Mol. Biochem. Parasitol.* **93**, 285–294 (1998).
40. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
41. van Lin, L. H. M., Pace, T., Janse, C. J., Scotti, R. & Ponzi, R. A long range restriction map of chromosomes 5 of *Plasmodium berghei* demonstates a chromosomes specific symmetrical subtelomeric organisation. *Mol. Biochem. Parasitol.* **86**, 111–115 (1997).
42. Janse, C. J., Ramesar, J., van den Berg, F. M. & Mons, B. *Plasmodium berghei*: *in vivo* generation and selection of karyotype mutants and non-gametocyte producer mutants. *Exp. Parasitol.* **74**, 1–10 (1992).
43. Liu, S. L. & Sanderson, K. E. Rearrangements in the genome of the bacterium *Salmonella typhi*. *Proc. Natl Acad. Sci. USA* **92**, 1018–1022 (1995).
44. Dame, J. B. & McCutchan, T. F. The four ribosomal DNA units of the malaria parasite *Plasmodium berghei*. Identification, restriction map and copy number analysis. *J. Biol. Chem.* **258**, 6984–6990 (1983).
45. van Lin, L. H. *et al.* Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. *Nucleic Acids Res.* **29**, 2059–2068 (2001).
46. Jareborg, N., Birney, E. & Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**, 815–824 (1999).
47. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373–376 (2001).
48. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
49. Carlton, J. M., Fidock, D. A., Djimde, A., Plowe, C. V. & Wellems, T. E. Conservation of a novel

vacuolar transporter in *Plasmodium* species and its central role in chloroquine resistance of *P. falciparum*. *Curr. Opin. Microbiol.* **4**, 415–420 (2001).
50. Weinbaum, F. I., Evans, C. B. & Tigelaar, R. E. An *in vitro* assay for T cell immunity to malaria in mice. *J. Immunol.* **116**, 1280–1283 (1976).
51. Landau, I. & Chabaud, A. G. Natural infection by 2 plasmodia of the rodent *Thamnomys rutilans* in the Central African Republic. *C.R. Acad. Sci. Hebd. Seances Acad. Sci. D* **261**, 230–232 (1965).
52. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
53. Beetsma, A. L., van de Wiel, T. J., Sauerwein, R. W. & Eling, W. M. *Plasmodium berghei* ANKA: purification of large numbers of infectious gametocytes. *Exp. Parasitol.* **88**, 69–72 (1998).
54. Gardner, M. J. *et al.* Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419**, 531–534 (2002).
55. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
56. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–43 (2001).
57. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
58. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
59. Ogurtsov, A. Y., Roytberg, M. A., Shabalina, S. A. & Kondrashov, A. S. OWEN: aligning long collinear regions of genomes. *Bioinformatics* (in the press).
60. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
61. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
62. Black, C. G., Wang, L., Hibbs, A. R., Werner, E. & Coppel, R. L. Identification of the *Plasmodium chabaudi* homologue of merozoite surface proteins 4 and 5 of *Plasmodium falciparum*. *Infect. Immun.* **67**, 2075–2081 (1999).

**Supplementary Information** accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

## Acknowledgements

We thank S. Cawley and T. Pace for collaborative work; J. Mendoza and J. Ramesar for technical support; C. Long for the gift of a *P. y. yoelii* cDNA library; R. Arcilla and W. Weiss for parasite material; and J. Eisen and S. Sullivan for critical reading of the manuscript. L.H.v.L. was supported by an INCO-DEV programme grant from the European Community; T.W.K. was supported by a Rijks Universiteit te Leiden studentship; J.D.R. was supported with funds from the Wellcome Trust. This project was funded by the US Department of Defense through cooperative agreement with the US Army Medical Research and Material Command and by the Naval Medical Research Center. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

## Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.C. (e-mail: carlton@tigr.org). Access to genome annotation data is available through the TIGR Eukaryotic Projects website (http://www.tigr.org) and PlasmoDB (http://www.plasmodb.org). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession number AABL00000000. The version described in this paper is the first version, AABL01000000.

articles

# A proteomic view of the *Plasmodium falciparum* life cycle

**Laurence Florens*, Michael P. Washburn†, J. Dale Raine‡, Robert M. Anthony§, Munira Grainger‖, J. David Haynes§¶, J. Kathleen Moch§, Nemone Muster*, John B. Sacci§#, David L. Tabb*☆, Adam A. Witney§#, Dirk Wolters†#, Yimin Wu**, Malcolm J. Gardner††, Anthony A. Holder‖, Robert E. Sinden‡, John R. Yates*† & Daniel J. Carucci§**

* *Department of Cell Biology, The Scripps Research Institute, SR-11, 10550 North Torrey Pines Road, La Jolla, California 92037, USA*
† *Department of Proteomics and Metabolomics, Torrey Mesa Research Institute, Syngenta Research & Technology, 3115 Merryfield Row, San Diego, California 92121-1125, USA*
‡ *Infection and Immunity Section, Department of Biological Sciences, Imperial College of Science, Technology & Medicine, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK*
§ *Naval Medical Research Center, Malaria Program (IDD), 503 Robert Grant Avenue, Room 3A40; and ¶ Department of Immunology, Walter Reed Army Institute of Research, Silver Spring, Maryland 20910-7500, USA*
‖ *The Division of Parasitology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*
☆ *Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA*
** *Malaria Research and Reference Reagent Resource Center, American Type Culture Collection, 10801 University Boulevard, Manassas, Virginia 20110-2209, USA*
†† *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*

**The completion of the *Plasmodium falciparum* clone 3D7 genome provides a basis on which to conduct comparative proteomics studies of this human pathogen. Here, we applied a high-throughput proteomics approach to identify new potential drug and vaccine targets and to better understand the biology of this complex protozoan parasite. We characterized four stages of the parasite life cycle (sporozoites, merozoites, trophozoites and gametocytes) by multidimensional protein identification technology. Functional profiling of over 2,400 proteins agreed with the physiology of each stage. Unexpectedly, the antigenically variant proteins of *var* and *rif* genes, defined as molecules on the surface of infected erythrocytes, were also largely expressed in sporozoites. The detection of chromosomal clusters encoding co-expressed proteins suggested a potential mechanism for controlling gene expression.**

The life cycle of *Plasmodium* is extraordinarily complex, requiring specialized protein expression for life in both invertebrate and vertebrate host environments, for intracellular and extracellular survival, for invasion of multiple cell types, and for evasion of host immune responses. Interventional strategies including anti-malarial vaccines and drugs will be most effective if targeted at specific parasite life stages and/or specific proteins expressed at these stages. The genomes of *P. falciparum*[1] and *P. yoelii yoelii*[2] are now completed and offer the promise of identifying new and effective drug and vaccine targets.

Functional genomics has fundamentally changed the traditional gene-by-gene approach of the pre-genomic era by capitalizing on the success of genome sequencing efforts. DNA microarrays have been successfully used to study differential gene expression in the abundant blood stages of the *Plasmodium* parasite[3,4]. However, transcriptional analysis by DNA microarrays generally requires microgram quantities of RNA and has been restricted to stages that can be cultivated *in vitro*, limiting current large-scale gene expression analyses to the blood stages of *P. falciparum*. As several key stages of the parasite life cycle, in particular the pre-erythrocytic stages, are not readily accessible to study, and as differential gene expression is in fact a surrogate for protein expression, global proteomic analyses offer a unique means of determining not only protein expression, but also subcellular localization and post-translational modifications.

We report here a comprehensive view of the protein complements isolated from sporozoites (the infectious form injected by the mosquito), merozoites (the invasive stage of the erythrocytes), trophozoites (the form multiplying in erythrocytes), and gametocytes (sexual stages) of the human malaria parasite *P. falciparum*. These proteomes were analysed by multidimensional protein identification technology (MudPIT), which combines in-line, high-resolution liquid chromatography and tandem mass spectrometry[5]. Two levels of control were implemented to differentiate parasite from host proteins. By using combined host–parasite sequence databases and noninfected controls, 2,415 parasite proteins were confidently identified out of thousands of host proteins; that is, 46% of all gene products were detected in four stages of the *Plasmodium* life cycle (Supplementary Table 1).

## Comparative proteomics throughout the life cycle

The sporozoite proteome appeared markedly different from the other stages (Table 1). Almost half (49%) of the sporozoite proteins

Table 1 **Comparative summary of the protein lists for each stage**

| Protein count | Sporozoites | Merozoites | Trophozoites | Gametocytes |
|---|---|---|---|---|
| 152 | X | X | X | X |
| 197 | – | X | X | X |
| 53 | X | – | X | X |
| 28 | X | X | – | X |
| 36 | X | X | X | – |
| 148 | – | – | X | X |
| 73 | – | X | – | X |
| 120 | X | – | – | X |
| 84 | – | X | X | – |
| 80 | X | – | X | – |
| 65 | X | X | – | – |
| 376 | – | – | – | X |
| 286 | – | – | X | – |
| 204 | – | X | – | – |
| 513 | X | – | – | – |
| 2,415 | 1,049 | 839 | 1,036 | 1,147 |

Whole-cell protein lysates were obtained from, on average, $17 \times 10^6$ sporozoites, $4.5 \times 10^9$ trophozoites, $2.75 \times 10^9$ merozoites, and $6.5 \times 10^9$ gametocytes.

# Present addresses: BRB 13-009, Department of Microbiology and Immunology, University of Maryland School of Medicine, 655 W. Baltimore St., Baltimore, Maryland 21201, USA (J.B.S.); Department of Medical Microbiology, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK (A.A.W.); and Ruhr-University Bochum, Institute of Analytical Chemistry, 44780 Bochum, Germany (D.W.).

were unique to this stage, which shared an average of 25% of its proteins with any other stage. On the other hand, trophozoites, merozoites and gametocytes had between 20% and 33% unique proteins, and they shared between 39% and 56% of their proteins. Consequently, only 152 proteins (6%) were common to all four stages. Those common proteins were mostly housekeeping proteins such as ribosomal proteins, transcription factors, histones and cytoskeletal proteins (Supplementary Table 1). Proteins were sorted into main functional classes based on the Munich Information Centre for Protein Sequences (MIPS) catalogue[6], with some adaptations for classes specific to the parasite, such as cell surface and apical organelle proteins (Fig. 1). When considering the annotated proteins in the database, some marked differences appeared between sporozoites and blood stages (Fig. 1). Although great care was taken to ensure that the results reflect the state of the parasite in the host, a portion of the data set may reflect the parasite's response to different purification treatments. However, the stage-specific detection of known protein markers at each stage established the relevance of our data set.

## The merozoite proteome

Merozoites are released from an infected erythrocyte, and after a short period in the plasma, bind to and invade new erythrocytes. Proteins on the surface and in the apical organelles of the merozoite mediate cell recognition and invasion in an active process involving an actin-myosin motor. Four putative components of the invasion motor[7], merozoite cap protein-1 (MCP1), actin, myosin A, and myosin A tail domain interacting protein (MTIP), were abundant merozoite proteins (Supplementary Table 2). Abundant merozoite surface proteins (MSPs) such as MSP1 and MSP2 are linked by a glycosylphosphatidyl (GPI) anchor to the membrane, and both have been implicated in immune evasion (reviewed in ref. 8). A second family of peripheral membrane proteins, represented by MSP3 and MSP6, was also detected (Fig. 2a), although these proteins are largely soluble proteins of the parasitophorous vacuole, which are released on schizont rupture. Other vacuolar proteins, such as the acidic basic repeat antigen (ABRA) and serine repeat antigen (SERA), were detected in the merozoite fraction, but some such as S-antigen[9] were not (Supplementary Table 2). Notably, MSP8 and a related MSP8-like protein were only identified in sporozoites (Fig. 2a). Some MSPs are diverse in sequence and may be extensively modified by proteolysis; these features, together with the association of a variety of peripheral and soluble proteins, provide for a complex surface architecture.

Many apical organellar proteins, in the micronemes and rhoptries, have a single transmembrane domain. Among these proteins, apical membrane antigen 1 (AMA1) and MAEBL were found in both sporozoite and merozoite preparations (Fig. 2a). Erythrocyte-binding antigens (EBA), such as EBA 175 and EBA 140/BAEBL, were found only in the merozoite and trophozoite fractions. Of note, the reticulocyte-binding protein (PfRH) family (PFD0110w, MAL13P1.176, PF13_01998, PFL2520w and PFD1150c), which has similarity with the Py235 family of *P. y. yoelii* rhoptry proteins and the *Plasmodium vivax* reticulocyte-binding proteins, was not detected in the merozoite fraction. Some PfRH proteins were, however, detected in sporozoites (Fig. 2a), including RH3, which is a transcribed pseudogene in blood stages[10]. Components of the low molecular mass rhoptry complex, the rhoptry-associated proteins (RAP) 1, 2 and 3, were all found in merozoites. RAP1 was also detected in sporozoites. The high molecular mass rhoptry protein complex (RhopH), together with ring-infected erythrocyte surface antigen (RESA), which is a component of dense granules, is transferred intact to new erythrocytes at or after invasion and may contribute to the host cell remodelling process. RhopH1, RhopH2 (PFI1445w; Ling, I. T., *et al.*, unpublished data) and RhopH3 were found in the merozoite proteome. RhopH1 (PFC0120w/PFC0110w) has been shown to be a member of the cyto-adherence linked asexual gene family (CLAG)[11]; however, the presence of CLAG9 in the merozoite fraction (Fig. 2a) suggests that CLAG9 may also be a RhopH protein, casting some doubt on the proposed role for this protein in cyto-adherence[12].

## The trophozoite proteome

After erythrocyte invasion the parasite modifies the host cell. The principal modifications during the initial trophozoite phase (lasting about 30 h) allow the parasite to transport molecules in and out of the cell, to prepare the surface of the red blood cell to mediate cyto-adherence, and to digest the cytoplasmic contents, particularly haemoglobin, in its food vacuole. In the next phase of schizogony (the final ~18 h of the asexual development in the blood cell), nuclear division is followed by merozoite formation and release.

Knob-associated histidine-rich protein (KAHRP) and erythrocyte membrane proteins 2 and 3 (EMP2 and -3) bind to the erythrocyte cytoskeleton (Fig. 2a). Of the proteins of the parasitophorous vacuole and the tubovesicular membrane structure extending into the cytoplasm of the red blood cell, three (the skeleton-binding protein 1, and exported proteins EXP1 and EXP2) were represented by peptides (Fig. 2a); although a fourth (Sar1 homologue, small GTP-binding protein; PFD0810w) was not. It is likely that one or more of the hypothetical proteins detected only in the trophozoite sample are involved in these unusual structures.

Digestion of haemoglobin is a major parasite catabolic process[13]. Members of the plasmepsin family (aspartic proteinases; PF14_0075 to PF14_0078)[14], falcipain family (cysteine proteinases; PF11_0161,

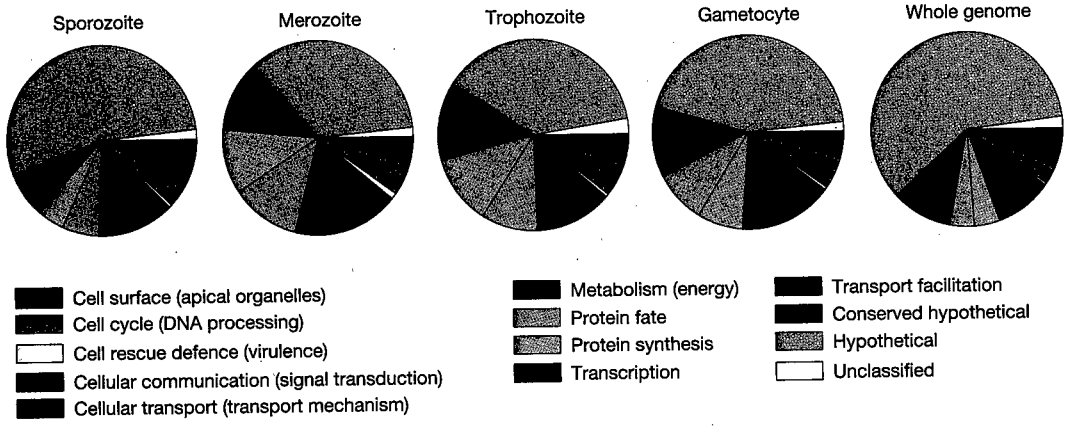



**Figure 1** Functional profiles of expressed proteins. Proteins identified in each stage are plotted as a function of their broad functional classification as defined by the MIPS catalogue[6]. To avoid redundancy, only one class was assigned per protein. The complete protein list is given in Supplementary Table 1.

PF11_0162 and PF11_0165)[15], and falcilysin (a metallopeptidase; PF13_0322)[16] implicated in this process were all clearly identified (Supplementary Table 1). Several proteases expressed in the merozoite and trophozoite fractions, and not involved in haemoglobin digestion, may be important in parasite release at the end of schizogony, invasion of the new cell, or merozoite protein processing. Possible candidates for this mechanism include cysteine proteinases of the falcipain and SERA families, or subtilisins such as SUB1 and SUB2, both located in apical organelles (Fig. 2a).

## The gametocyte proteome

Stage V gametocytes are dimorphic, with a male:female ratio of 1:4. They are arrested in the cell cycle until they enter the mosquito where development is induced within minutes to form the male and female gametes. Gametocyte structure reflects these ensuing fates; that is, the female has abundant ribosomes and endoplasmic reticulum/vesicular network to re-initiate translation, whereas the male is largely devoid of ribosomes and is terminally differentiated[17].

Gametocyte-specific transcription factors, RNA-binding proteins, and gametocyte-specific proteins involved in the regulation of messenger RNA processing (particularly splicing factors, RNA helicases, RNA-binding proteins, ribonucleoproteins (RNPs) and small nuclear ribonucleoprotein particles (snRNPS)) were highly represented in the gametocyte proteome (Supplementary Table 1). Transcription in the terminally differentiated gametocytes is 'suppressed', but the female gametocytes contain mRNAs encoding gamete/zygote/ookinete surface antigens (for example, P25/28)

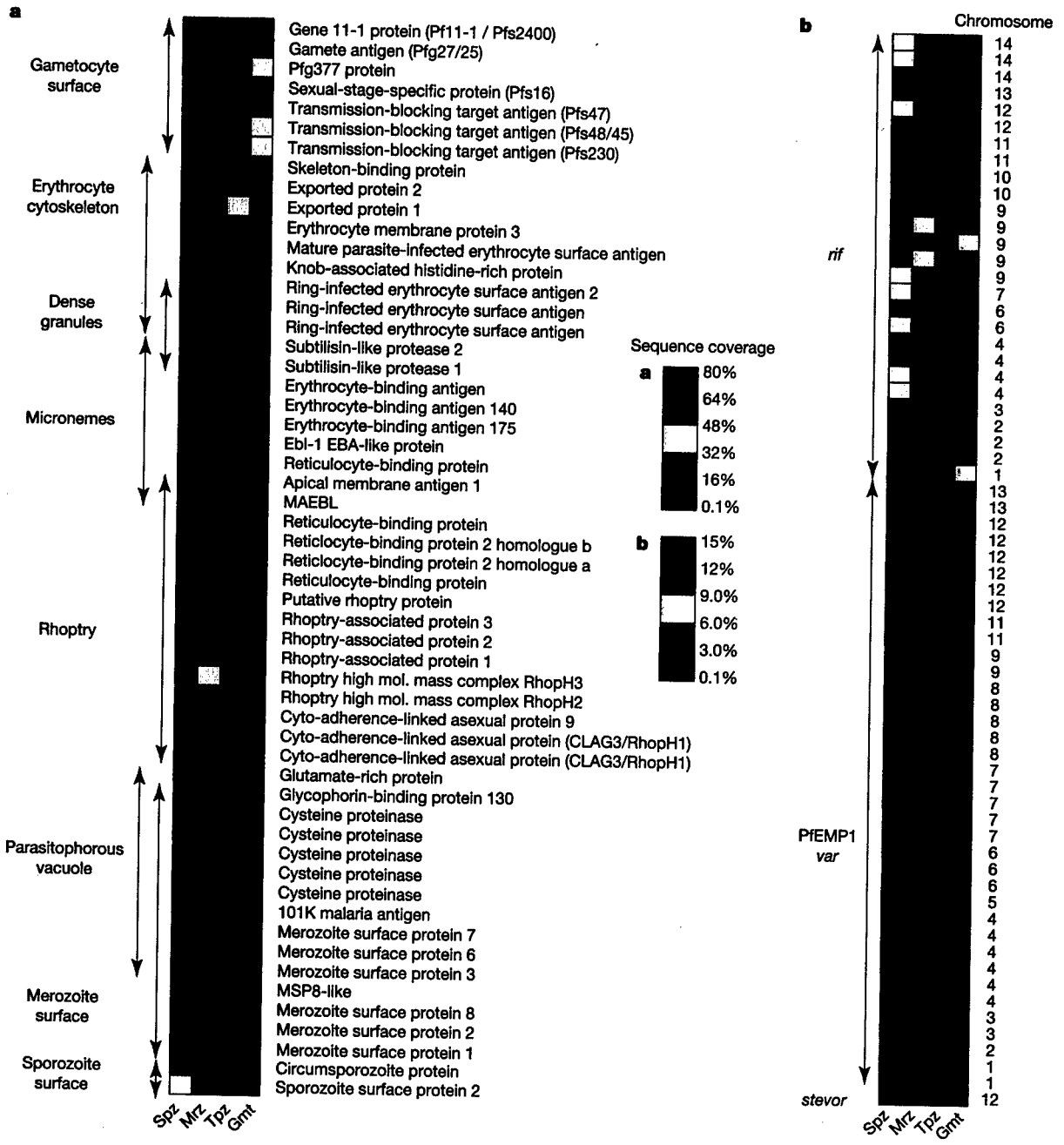



**Figure 2** Expression patterns of known stage-specific proteins. **a**, Cell surface, organelle, and secreted proteins are plotted as a function of their known subcellular localization. **b**, *stevor*, *var* and *rif* polymorphic surface variants are plotted as a function of the chromosome encoding their genes. The matrices are colour-coded by sequence coverage measured in each stage (proteins not detected in a stage are represented by black squares). Locus names associated with these proteins are listed in Supplementary Table 2. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.

that are subject to post-transcriptional control; this control is released rapidly during gamete development[17]. Ribosomal proteins were largely represented: 82% of known small subunit (SSU) proteins and 69% of known large subunit (LSU) proteins were detected in gametocytes compared to 94% and 82%, respectively, from all stages examined (Supplementary Table 1). We suggest that this reflects the accumulation of ribosomes in the female gametocyte to accommodate for the sudden increase in protein synthesis required during gametogenesis and early zygote development.

Other protein groupings highly represented in the gametocyte were in the cell cycle/DNA processing and energy classes (Fig. 1). The former is consistent with the biological observation that the mature gametocyte is arrested in G0 of the cell cycle and will require a full complement of pre-existing cell cycle regulatory cascades to respond, within seconds, to the gametogenesis stimuli (that is, xanthurenic acid and a drop in temperature)[18]. Metabolic pathways of the malaria parasite may be stage-specific, with asexual blood stage parasites dependent on glycolysis and conversion of pyruvate to lactate (L-lactate dehydrogenase) for energy. In the gametocyte and sporozoite preparations, peptides from enzymes involved in the mitochondrial tricarboxylic acid (TCA) cycle and oxidative phosphorylation were identified (Table 2). This observation suggests that gametocytes have fully functional mitochondria as a pre-adaptation to life in the mosquito, as suggested by morphological and biochemical studies[19] and their sensitivity to anti-malarials attacking respiration (primaquine and artimesinin-based products)[17]. It will be interesting to observe whether other mosquito and liver stages, which show similar drug sensitivities, express the same metabolic proteome.

Cell surface proteins (Fig. 1) included most of the known surface antigens (Fig. 2a and Supplementary Table 2). However, Pfs35 and a sexual stage-specific kinase (PF13_0258) were not detected. Nevertheless the cultured gametocytes analysed in this study expressed a specific repertoire of rifin and PfEMP1 proteins (Fig. 2b and Supplementary Table 2). Together these observations suggest that the gametocyte, which is very long-lived in the red blood cell (that is, 9–12 days compared with 2 days for the pathogenic asexual parasites), expresses a limited repertoire of the highly polymorphic families of surface antigens so widely represented in the asexual parasites.

## The sporozoite proteome

Sporozoites are injected by the mosquito during ingestion of a blood meal. Although, they are in the blood stream for only minutes, sporozoites probably require mechanisms to evade the host humoral immune system in order for at least a fraction of the thousands of sporozoites injected by the mosquito to survive the hostile environment in the blood and successfully invade hepatocytes.

The main class of annotated sporozoite proteins identified was cell surface and organelle proteins (Fig. 1). Sporozoites are an invasive stage and possess the apical complex machinery involved in host cell invasion. As observed in the analysis of the *P. y. yoelii* sporozoite transcriptome[20], actin and myosin were found in the motile sporozoites (Supplementary Table 2). Many proteins associated with rhoptry, micronemes and dense granules were detected (Fig. 2a). Among the proteins found were known markers of the sporozoite stage, such as the circumsporozoite protein (CSP) and sporozoite surface protein 2 (SSP2; also known as TRAP), both present in large quantities at the sporozoite surface (Fig. 2a). Peptides derived from CTRP (circumsporozoite protein and thrombospondin-related adhesive protein (TRAP)-related protein), an ookinete cell surface protein involved in recognition and/or motility[21], were detected in the sporozoite fractions (Supplementary Table 1).

Most surprisingly, peptides derived from multiple *var* (coding for PfEMP1) and *rif* genes were identified in the sporozoite samples. PfEMP1 and rifins are coded for by large multigene families (*var* and *rif*)[22,23] and are present on the surface of the infected red blood cell. No peptides derived from *rif* genes were identified in the trophozoite sample, whereas sporozoites expressed 21 different rifins and 25 PfEMP1 isoforms (Fig. 2b); that is, a total of 14% of the *rif* genes and 33% of the *var* genes encoded by the genome. Furthermore, very little overlap was observed between stages: only ten PfEMP1 and two rifin isoforms expressed in sporozoites were found in other stages. Whereas in the blood stream the asexual stage parasites undergo asexual multiplication and therefore have an opportunity to undergo antigenic 'switching' of the variant antigen genes, the non-replicative sporozoites may not have this opportunity. Expressing such a polymorphic array of *var* (PfEMP1) and *rif* genes could be part of a sporozoite survival mechanism.

## Chromosomal clusters encoding co-expressed proteins

The distinct proteomes of each stage of the *Plasmodium* life cycle suggested that there is a highly coordinated expression of *Plasmodium* genes involved in common processes. Co-expression groups are a widespread phenomenon in eukaryotes, where mRNA array analyses have been used to establish gene expression profiles. Analysis of co-regulated gene groups facilitates both searching for regulatory motifs common to co-regulated genes, and predicting protein function on the basis of the 'guilt by association' model. Furthermore, mRNA analyses in *Saccharomyces cerevisiae*[24] and *Homo sapiens*[25,26] have demonstrated that co-regulated genes do not map to random locations in the genome but are in fact

Table 2 **Examples on enzymes in stage-specific metabolic pathways**

| Locus | Stage | | | | Enzyme | EC number† | Reaction catalysed |
|---|---|---|---|---|---|---|---|
| | Spz* | Mrz* | Tpz* | Gmt* | | | |
| End of glycolysis | | | | | | | |
| PF10_0363 | 1.2 | – | 2.4 | – | Pyruvate kinase | 2.7.1.40 | *P*-enolpyruvate to pyruvate |
| MAL6P1.160 | 8.6 | 66.9 | 18.8 | 14.7 | Pyruvate kinase | | |
| PF13_0141 | 46.2 | 83.9 | 70.9 | 78.8 | L-lactate dehydrogenase | 1.1.1.27 | Pyruvate to lactate |
| TCA cycle and oxidative phosphorylation | | | | | | | |
| PF10_0218 | 12.3 | – | – | – | Citrate synthase | 4.1.3.7 | Acetyl coA + oxaloacetate to citrate |
| PF13_0242 | 3.2 | – | 16.9 | 8.8 | Isocitrate dehydrogenase (NADP) | 1.1.1.41 | Isocitrate to 2-oxoglutarate + $CO_2$ |
| PF08_0045 | 2.9 | – | 2.2 | 23.1 | 2-Oxoglutarate dehydrogenase e1 component | 1.2.4.2 | 2-Oxoglutarate to succinyl CoA |
| PF10_0334 | – | – | 3.5 | 27.7 | Flavoprotein subunit of succinate dehydrogenase | 1.3.5.1 | Succinate to fumarate |
| PFL0630w | 3.7 | – | – | 12.1 | Iron-sulphur subunit of succinate dehydrogenase | | |
| PF14_0373 | – | – | – | 12.7 | Ubiquinol cytochrome oxidoreductase | 1.10.2.2. | Ubiquinol to cytochrome *c* reductase in electron transport |
| PFB0795w | – | – | – | 14.2 | ATP synthase F1, α-subunit | | |
| PFI1365w | – | – | – | 8.8 | Cytochrome *c* oxidase subunit | 1.9.3.1 | |
| PFI1340w | – | – | – | 8.8 | Fumarate hydratase | 4.2.1.2 | Fumarate to malate |
| MAL6P1.242 | 30.4 | – | – | 40.9 | Malate dehydrogenase | 1.1.1.37 | Malate to oxaloacetate |

*Plasmodium* metabolic pathways can be found at http://www.sites.huji.ac.il/malaria/. Spz, sporozoite; mrz, merozoite; tpz, trophozoite; gmt, gametocyte.
*The sequence coverage (that is, the percentage of the protein sequence covered by identified peptides) measured in each stage is reported.
†Enzyme Commission (EC) numbers are reported for each protein.

frequently organized into gene clusters on a chromosome. Gene clustering in *Plasmodium* species has been demonstrated. Ordered arrays of genes involved in virulence and antigenic variation (for example, *var*, *vir* and *rif* genes) are located in the subtelomeric regions of the chromosomes[27,28].

To determine whether gene clustering exists along the entire *P. falciparum* genome, genes whose protein products were detected in our analysis were mapped onto all 14 chromosomes in a stage-dependent manner (Fig. 3a). The 2,415 proteins identified represented an average of 45% of the open reading frames (ORFs) predicted per chromosome. The number of protein hits by chromosome was similar for all stages: sporozoite, merozoite, trophozoite and gametocyte protein lists constituting 19.7%, 15.8%, 19.5% and 21.6% of the predicted ORFs per chromosomes, respectively. Groups of three or more consecutive loci whose protein products were detected in a particular stage were defined as chromosomal clusters encoding co-expressed proteins (Fig. 3b). On the basis of this definition a total of 98 clusters containing 3 loci, 32 clusters containing 4 loci, 5 clusters containing 5 loci, and 3 clusters containing 6 loci were identified (Supplementary Table 3). For each chromosome, the frequency of finding clusters encoding co-expressed proteins containing 3–6 adjacent loci markedly exceeded the probability of finding such clusters by chance (see the footnote of Supplementary Table 3 for details on the probability calculation). Therefore, chromosomal clusters encoding co-expressed proteins were prevalent in the *P. falciparum* genome.

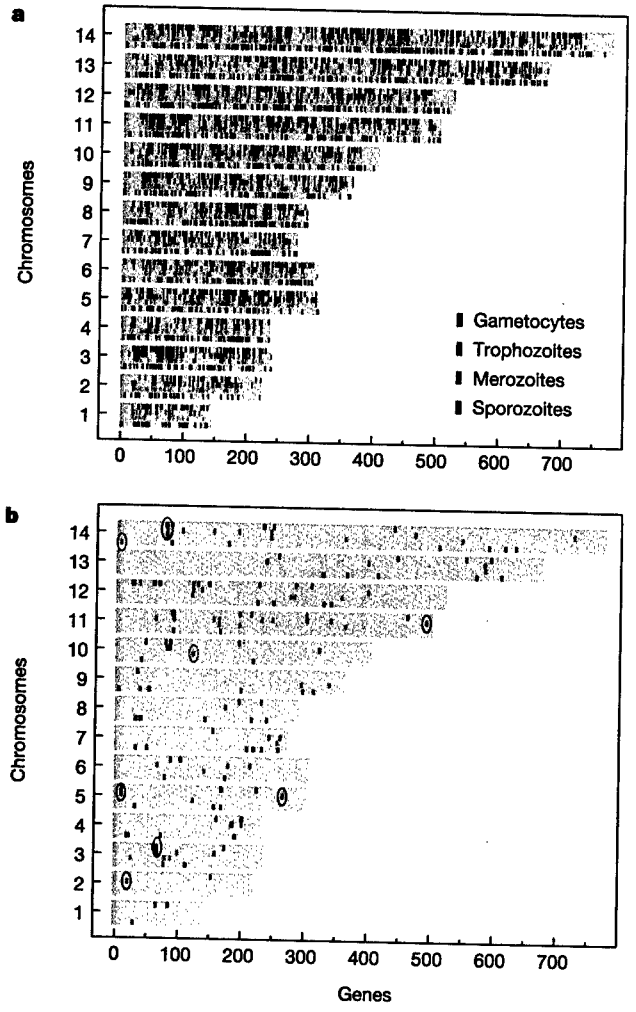



**Figure 3** Distribution of expressed proteins by chromosome. **a**, For each stage, genes whose products were detected (coloured vertical bars) are plotted in the order they appear on their chromosome (grey boxes). **b**, Groups of at least three consecutive expressed genes are defined as chromosomal clusters of co-expressed proteins. Examples of such clusters, circled in **b**, are specified in Table 3 and the complete description of the 138 clusters can be found in Supplementary Table 3.

Functionally related genes have been shown to cluster in the *S. cerevisiae*[24] and human genomes[26]. This phenomenon also occurs in *P. falciparum*. A total of 138 clusters encoding co-expressed proteins were identified and 67 of them (49%) contained at least two loci that have been functionally annotated. Of these 67 clusters, 30 contained at least two loci whose annotation clearly indicates that the proteins are functionally related. For example, clusters on chromosomes 3, 5 and 10 contained ribosomal proteins, proteins involved in protein modification, and proteins involved in nucleotide metabolism, respectively (Table 3). Chromosome 14 contained a cluster of four aspartic proteases co-expressed in all of the blood stages (Table 3). This cluster was not detected in sporozoites, where no haemoglobin degradation is expected to occur. Interestingly, whereas the falcipain gene cluster on chromosome 11 appeared in our analysis as a cluster of co-expressed proteins (Supplementary Table 3), the SERA gene cluster on chromosome 2, coding for proteins that share a papain-like sequence motif[29], did not. Of the ten sporozoite-specific clusters, five involved *var* and *rif* genes, such as the *rif* cluster located in the subtelomeric domain of chromosome 14 (Table 3). On the basis of their presence in clusters encoding co-expressed proteins, we were able to suggest functional roles for 24 proteins annotated as hypothetical in the *P. falciparum* genome (Supplementary Table 3). For example, a gametocyte-specific cluster on chromosome 13 encoded two transmission-blocking antigens (Pfs48/45 and Pfs47) and a hypothetical protein, PF13_0246, which might be a gametocyte surface protein. Two clusters on chromosomes 2 and 11 were highly specific to the trophozoite stage (Table 3). Each of these clusters contained well-known secreted and surface proteins, namely KAHRP, PfEMP3, antigen 332, and RESA, all of which have been implicated in knob formation. The highly coordinated expression of these genes makes the three hypothetical proteins listed in these trophozoite-specific gene clusters possible candidates for involvement in cyto-adherence.

## Discussion

Although sample handling is a principal consideration when studying pathogens, the expression of large numbers of previously identified proteins was consistent with their published expression profiles, validating our data set as a meaningful sampling of each stage's proteome. This is a particularly important aspect of our analysis as 65% of the 5,276 genes encoded by the *P. falciparum* genome are annotated as hypothetical[1], and of the 2,415 expressed proteins we identified, 51% are hypothetical proteins (Supplementary Table 1). Our results confirmed that these hypothetical ORFs predicted by gene modelling algorithms were indeed coding regions. Furthermore, from all four stages analysed, we identified 439 proteins predicted to have at least one transmembrane segment or a GPI addition signal (18% of the data set) and 304 soluble proteins with a signal sequence; that is, potentially secreted or located to organelles. Well over half of the secreted proteins and integral membrane proteins detected were annotated as hypothetical (Supplementary Table 4). The obvious interest in this class of proteins is that, with no homology to known proteins, they represent potential *Plasmodium*-specific proteins and may provide targets for new drug and vaccine development.

Our comprehensive large-scale analysis of protein expression showed that most surface proteins are more widely expressed than initially thought. In particular, the *var* and *rif* genes, which were thought to be involved in immune evasion only in the blood stage, have now been shown to be expressed in apparently large and varied numbers at the sporozoite stage. These surface proteins might be involved in general interaction processes with host cells and/or immune evasion. An alternative hypothesis is that stage-specific

Table 3 **Examples of chromosomal gene clusters encoding co-expressed proteins**

| Chromosome | ID | Locus | Stage | | | | Description | Class | SP | TM |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Spz | Mrz | Tpz | Gmt | | | | |
| 3 | 64 | PFC0285c | 2.1 | 12.7 | 33.2 | 18.7 | T-complex protein β-subunit | Protein fate | 0 | 0 |
| 3 | 65 | PFC0290w | 8.3 | – | 33.8 | 18.6 | 40S ribosomal protein S23 | Protein synthesis | 0 | 0 |
| 3 | 66 | PFC0295c | – | 14.9 | 52.5 | 21.3 | 40S ribosomal protein S12 | Protein synthesis | 0 | 0 |
| 3 | 67 | PFC0300c | – | 12.1 | 30.4 | 17.9 | 60S ribosomal protein L7 | Protein synthesis | 0 | 0 |
| 5 | 263 | PFE1345c | – | – | 1.9 | 1.6 | Minichromosome maintenance protein 3 | Cell transport | 0 | 0 |
| 5 | 264 | PFE1350c | – | – | 22.4 | – | Ubiquitin-conjugating enzyme | Protein fate | 0 | 0 |
| 5 | 265 | PFE1355 | – | 4.8 | 2.6 | 2.6 | Ubiquitin carboxy-terminal hydrolase | Protein fate | 0 | 0 |
| 5 | 266 | PFE1360c | – | – | 7.7 | – | Methionine aminopeptidase | Protein fate | 0 | 0 |
| 10 | 119 | PF10_0121 | 10.8 | 74.5 | 29 | – | Hypoxanthine phosphoribosyltransferase | Metabolism | 0 | 0 |
| 10 | 120 | PF10_0122 | 5.4 | 6.1 | – | 6.1 | Phosphoglucomutase | Metabolism | 0 | 0 |
| 10 | 121 | PF10_0123 | – | 11.7 | – | – | GMP synthetase | Metabolism | 0 | 0 |
| 10 | 122 | PF10_0124 | 0.9 | 1.8 | – | – | Hypothetical protein | | 0 | 0 |
| 14 | 74 | PF14_0074 | 26.6 | – | – | 4.9 | Hypothetical protein | | 0 | 0 |
| 14 | 75 | PF14_0075 | – | 26.5 | 43.2 | 47.4 | Plasmepsin | Protein fate | 1 | 0 |
| 14 | 76 | PF14_0076 | – | 6.6 | 35.2 | 10 | Plasmepsin 1 | Protein fate | 1 | 0 |
| 14 | 77 | PF14_0077 | – | 21.2 | 43 | 11.5 | Plasmepsin 2 | Protein fate | 1 | 0 |
| 14 | 78 | PF14_0078 | – | 14.2 | 52.8 | 29.9 | HAP protein | Protein fate | 1 | 0 |
| 14 | 2 | PF14_0002 | 3.5 | – | – | – | Rifin | Surface or organelles | 0 | 1 |
| 14 | 3 | PF14_0003 | 7.9 | – | – | – | Rifin | Surface or organelles | 1 | 2 |
| 14 | 4 | PF14_0004 | 6.5 | – | – | – | Rifin | Surface or organelles | 1 | 2 |
| 2 | 18 | PFB0090c | – | – | 3 | – | Hypothetical protein, conserved | | 0 | 0 |
| 2 | 19 | PFB0095c | – | – | 3.4 | – | Erythrocyte membrane protein 3 | Surface or organelles | 1 | 0 |
| 2 | 20 | PFB0100c | – | 1.5 | 24.8 | – | Knob-associated histidine-rich protein | Surface or organelles | 1 | 0 |
| 11 | 489 | PF11_0506 | – | – | 6.3 | 4.4 | Hypothetical protein | | 0 | 1 |
| 11 | 490 | PF11_0507 | – | – | 0.8 | – | Antigen 332 | Surface or organelles | 0 | 0 |
| 11 | 491 | PF11_0508 | – | – | 3.3 | – | Hypothetical protein | | 0 | 0 |
| 11 | 492 | PF11_0509 | – | 6.4 | 3 | – | RESA | Surface or organelles | 0 | 0 |
| 13 | 443 | PF13_0246 | 4.5 | – | – | 8.6 | Hypothetical protein | | 0 | 0 |
| 13 | 444 | PF13_0247 | – | – | – | 32.4 | Transmission-blocking target antigen precursor (Pfs48/45) | Surface or organelles | 1 | 1 |
| 13 | 445 | PF13_0248 | – | – | – | 7.1 | Transmission-blocking target antigen precursor (Pfs47) | Surface or organelles | 1 | 1 |

Clusters of at least three consecutive genes encoding co-expressed proteins are reported with their position (ID) on the chromosome, the sequence coverage measured for these proteins in each stage (%), their current annotation and functional class, and the predicted presence of signal peptide (SP) or transmembrane domains (TM) (based on the TMHMM[43], a transmembrane (TM) helices prediction method based on a hidden Markov model (HMM), big-PI Predictor[44] and SignalP[45] algorithms).

regulation is not as exact as previously thought.

One mechanism of protein expression control that contributes to stage specificity in *P. falciparum* arises from the chromosomal clustering of genes encoding co-expressed proteins. The clusters described in this study demonstrate a widespread high order of chromosomal organization in *P. falciparum* and probably correspond to regions of open chromatin allowing for co-regulated gene expression. The high (A + T) content of the *P. falciparum* genome makes the identification of regulatory sequences such as promoters and enhancers challenging[31,32]. Focusing analyses on stage-specific and multi-stage clusters will facilitate finding stage-specific and general *cis*-acting sequences in the *Plasmodium* genome and will help decipher gene expression regulation during the parasite life cycle.

The malaria parasite is a complex multi-stage organism, which has co-evolved in mosquitoes and vertebrates for millions of years. Designing drugs or vaccines that substantially and persistently interrupt the life cycle of this complex parasite will require a comprehensive understanding of its biology. The *P. falciparum* genome sequence and comparative proteomics approaches may initiate new strategies for controlling the devastating disease caused by this parasite. □

## Methods

### Parasite material

*Plasmodium falciparum* clone 3D7 (Oxford) was used throughout. Sporozoites were initially isolated from the salivary glands of *Anopheles stephansi* mosquitoes, 14 days after infection, by centrifugation in a Renograffin 60 gradient, as described[33]. Four sporozoite samples were used as is. A fifth sample underwent an additional purification step on Dynabeads M-450 Epoxy coupled to NFS1 (an anti-*P. falciparum* CS protein monoclonal antibody)[34] according to the manufacturer's instructions (Dynal). Trophozoite-infected erythrocytes from synchronized cultures were purified on 70% Percoll-alanine[30], and the trophozoites released from the erythrocytes[35]. Of the of 260 parasitized erythrocytes counted by Giemsa-stained thin-blood film, 100% were identified as trophozoites. Merozoites were prepared essentially as described in ref. 36, using highly synchronized schizonts and purifying the merozoites by passage through membrane filters. Starting with synchronized asexual parasites grown in suspension culture as described[37,38], gametocytes were prepared by daily media changes of static cultures at 37 °C. When there were very few mature asexual stages present, gametocyte-infected erythrocytes were collected from the 52.5%/45% and 45%/30% interfaces of a Percoll gradient[39]. The gametocytes consisted mostly of stage IV and V parasites with minor contamination (<3%) from mixed asexual stage parasites. Finally, cellular debris from the upper bodies of parasite-free *A. stephansi* and non-infected human erythrocytes were used as controls for sporozoites and blood-stage parasites, respectively. Every effort was made to minimize enzymatic activity and protein degradation during sampling, and the subsequent isolation of the parasites; however, we cannot exclude that some of the differences in protein profiles that we observe between the different life-cycle stages may be a consequence of the sample-handling procedures.

### Cell lysis

Five sporozoite, four merozoite, four trophozoite and three gametocyte preparations were lysed, digested and analysed independently. Cell pellets were first diluted ten times in 100 mM Tris-HCl pH 8.5, and incubated in ice for 1 h. After centrifugation at 18,000 *g* for 30 min, supernatants were set aside and microsomal membrane pellets were washed in 0.1 M sodium carbonate, pH 11.6. Soluble and insoluble protein fractions were separated by centrifugation at 18,000 *g* for 30 min. Supernatants obtained from both centrifugation steps were either combined (sporozoites, trophozoites and merozoites) or digested and analysed independently (gametocytes).

### Peptide generation and analysis

The method follows that of Washburn *et al.*[5], with the exception that Tris(2-carboxyethyl)phosphine hydrochloride (TCEP-HCl; Pierce) was used to reduce urea-denatured proteins. Peptide mixtures were analysed through MudPIT as described[5].

### Protein sequence databases

The *P. falciparum* database contained 5,283 protein sequences. Spectra resulting from contaminant mosquito and erythrocyte peptides had to be taken into account in the sporozoite and blood-stage samples, respectively. Tandem mass spectrometry (MS/MS) data sets from blood stages were therefore searched against a database containing both *P. falciparum* protein sequences and 24,006 ORFs from the human, mouse and rat RefSeq NCBI databases. At the date of the searches, the *Anopheles gambiae* genome was not available. The NCBI database contained 922 *Anopheles* and 313 *Aedes* proteins, which were combined to the 14,335 ORFs of the NCBI *Drosophila melanogaster*[40] database to create a control diptera database. Finally, these databases were complemented with a set of 172 known protein contaminants, such as proteases, bovine serum albumin and human keratins.

## MS/MS data set analysis

The SEQUEST algorithm was used to match MS/MS spectra to peptides in the sequence databases[41]. To account for carboxyamidomethylation, MS/MS data sets were searched with a relative molecular mass of 57,000 ($M_r$ 57K) added to the average molecular mass of cysteines. Peptide hits were filtered and sorted with DTASelect[42]. Spectra/peptide matches were only retained if they were at least half-tryptic (Lys or Arg at either end of the identified peptide) and with minimum cross-correlation scores (XCorr) of 1.8 for +1, 2.5 for +2, and 3.5 for +3 spectra and DeltaCn (top match's XCorr minus the second-best match's XCorr divided by the top match's XCorr) of 0.08. Peptide hits were deemed unambiguous only if they were not found in non-infected controls and were uniquely assigned to parasite proteins by searching against combined parasite–host databases. Finally, for low coverage loci, peptide/spectrum matches were visually assessed on two main criteria: any given MS/MS spectrum had to be clearly above the baseline noise, and both *b* and *y* ion series had to show continuity. The Contrast tool[42] was used to compare and merge protein lists from replicate sample runs and to compare the proteomes established for the four stages.

Received 31 July; accepted 9 September 2002; doi:10.1038/nature01107.


1. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
2. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419,** 512–519 (2002).
3. Ben Mamoun, C. *et al.* Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol. Microbiol.* **39,** 26–36 (2001).
4. Hayward, R. E. *et al.* Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* **35,** 6–14 (2000).
5. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19,** 242–247 (2001).
6. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34 (2002).
7. Pinder, J. C. *et al.* Actomyosin motor in the merozoite of the malaria parasite, *Plasmodium falciparum*: implications for red cell invasion. *J. Cell Sci.* **111,** 1831–1839 (1998).
8. Holder, A. A. *Malaria Vaccine Development: a Multi-immune Response and Multi-stage Perspective* (ed. Hoffman, S. L.) 77–104 (ASM Press, Washington, 1996).
9. Coppel, R. L. *et al.* Isolate-specific S-antigen of *Plasmodium falciparum* contains a repeated sequence of eleven amino acids. *Nature* **306,** 751–756 (1983).
10. Taylor, H. M. *et al. Plasmodium falciparum* homologue of the genes for *Plasmodium vivax* and *Plasmodium yoelii* adhesive proteins, which is transcribed but not translated. *Infect. Immun.* **69,** 3635–3645 (2001).
11. Kaneko, O. *et al.* The high molecular mass rhoptry protein, RhopH1, is encoded by members of the clag multigene family in *Plasmodium falciparum* and *Plasmodium yoelii*. *Mol. Biochem. Parasitol.* **118,** 223–231 (2001).
12. Trenholme, K. R. *et al.* clag9: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc. Natl Acad. Sci. USA* **97,** 4029–4033 (2000).
13. Klemba, M. & Goldberg, D. E. Biological roles of proteases in parasitic protozoa. *Annu. Rev. Biochem.* **71,** 275–305 (2002).
14. Banerjee, R. *et al.* Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proc. Natl Acad. Sci. USA* **99,** 990–995 (2002).
15. Rosenthal, P. J., Sijwali, P. S., Singh, A. & Shenai, B. R. Cysteine proteases of malaria parasites: targets for chemotherapy. *Curr. Pharm. Des.* **8,** 1659–1672 (2002).
16. Eggleson, K. K., Duffin, K. L. & Goldberg, D. E. Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **274,** 32411–32417 (1999).
17. Sinden, R. E., Butcher, G. A., Billker, O. & Fleck, S. L. Regulation of infectivity of *Plasmodium* to the mosquito vector. *Adv. Parasitol.* **38,** 53–117 (1996).
18. Billker, O., Shaw, M. K., Margo, G. & Sinden, R. E. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature* **392,** 289–292 (1998).
19. Krungkrai, J., Prapunwattana, P. & Krungkrai, S. R. Ultrastructure and function of mitochondria in gametocytic stage of *Plasmodium falciparum*. *Parasite* **7,** 19–26 (2000).
20. Kappe, S. H. *et al.* Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl Acad. Sci. USA* **98,** 9895–9900 (2001).
21. Dessens, J. T. *et al.* CTRP is essential for mosquito infection by malaria ookinetes. *EMBO J.* **18,** 6221–6227 (1999).
22. Deitsch, K. W. & Wellems, T. E. Membrane modifications in erythrocytes parasitized by *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **76,** 1–10 (1996).
23. Kyes, S. A., Rowe, J. A., Kriek, N. & Newbold, C. I. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **96,** 9333–9338 (1999).
24. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26,** 183–186 (2000).
25. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289–1292 (2001).
26. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.* **31,** 180–183 (2002).
27. Hernandez-Rivas, R. *et al.* Expressed var genes are found in *Plasmodium falciparum* subtelomeric regions. *Mol. Cell Biol.* **17,** 604–611 (1997).
28. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410,** 839–842 (2001).
29. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
30. Kanaani, J. & Ginsburg, H. Metabolic interconnection between the human malarial parasite *Plasmodium falciparum* and its host erythrocyte. *J. Biol. Chem.* **264,** 3194–3199 (1989).
31. Dechering, K. J. *et al.* Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell Biol.* **19,** 967–978 (1999).
32. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
33. Pacheco, N. D., Strome, C. P., Mitchell, F., Bawden, M. P. & Beaudoin, R. L. Rapid, large-scale isolation of *Plasmodium berghei* sporozoites from infected mosquitoes. *J. Parasitol.* **65,** 414–417 (1979).
34. Mellouk, S. *et al.* Evaluation of an *in vitro* assay aimed at measuring protective antibodies against sporozoites. *Bull. World Health Organ.* **68** Suppl., 52–59 (1990).
35. Rabilloud, T. *et al.* Analysis of membrane proteins by two-dimensional electrophoresis: comparison of the proteins extracted from normal or *Plasmodium falciparum*-infected erythrocyte ghosts. *Electrophoresis* **20,** 3603–3610 (1999).
36. Blackman, M. J. Purification of *Plasmodium falciparum* merozoites for analysis of the processing of merozoite surface protein-1. *Methods Cell Biol.* **45,** 213–220 (1994).
37. Haynes, J. D. & Moch, J. K. Automated synchronization of *Plasmodium falciparum* parasites by culture in a temperature-cycling incubator. *Methods Mol. Med.* **72,** 489–497 (2002).
38. Haynes, J. D., Moch, J. K. & Smoot, D. S. Erythrocytic malaria growth or invasion inhibition assays with emphasis on suspension culture GIA. *Methods Mol. Med.* **72,** 535–554 (2002).
39. Carter, R., Ranford-Cartwright, L. & Alano, P. The culture and preparation of gametocytes of *Plasmodium falciparum* for immunochemical, molecular, and mosquito infectivity studies. *Methods Mol. Biol.* **21,** 67–88 (1993).
40. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
41. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
42. Tabb, D. L., McDonald, W. H. & Yates, J. R. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26 (2002).
43. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
44. Eisenhaber, B., Bork, P. & Eisenhaber, F. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* **11,** 1155–1161 (1998).
45. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6 (1997).


**Supplementary Information** accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

## Acknowledgements

We are grateful to J. Graumann, R. Sadygov, G. Chukkapalli, A. Majumdar and R. Sinkovits for computer programming; C. Deciu for the probability calculations; and C. Delahunty and C. Vieille for critical reading of the manuscript. The authors acknowledge the support of the Office of Naval Research, the US Army Medical Research and Material Command, and the National Institutes of Health (to J.R.Y.). J.D.R. is funded by a Wellcome Trust Prize Studentship. We thank the scientists and funding agencies comprising the international Malaria Genome Project for making sequence data from the genome of *P. falciparum* clone 3D7 public before publication of the completed sequence. The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

## Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.Y. (e-mail: jyates@scripps.edu).